PAST, logiciel statistique naturaliste

Pierre DIEUMEGARD
professeur de SVT
Lycée Pothier
45044 Orléans

courriel : pierre.dieumegard@ac-orleans-tours.fr

Introduction : PAST, un logiciel de statistiques pour les sciences de la nature

« PAST » signifie « PAlaeontological STatistics », c'est à dire « Statistiques pour la paléontologie ». C'est un logiciel dont l'auteur principal est Øyvind Hammer, du Museum d'Histoire Naturelle de l'Université d'Oslo (Norvège). On peut le télécharger sur le site http://folk.uio.no/ohammer/past/

La paléontologie, qui est l'étude des fossiles, est souvent considérée à tort comme une vieille science, sous prétexte que son objet est l'étude des vieilles choses. Beaucoup imaginent que ses pratiquants récupèrent des fossiles dans des carrières et les mettent dans des tiroirs avec une étiquette en latin. Ils n'imaginent pas que ce puisse être une science utilisant des outils et des méthodes modernes.

C'est une erreur. C'est justement parce que la paléontologie étudie des objets anciens, déformés, disparates, avec souvent des données manquantes, des portions disparues... qu'elle a besoin d'outils scientifiques performants, en particulier dans le domaine statistique.

La paléontologie est la science qui étudie les squelettes, les coquilles, les traces des organismes ayant vécu il y a longtemps (plusieurs milliers, millions ou milliards d'années). Elle est donc à l'intersection de la biologie et de la géologie. En dehors de la paléontologie, les statistiques de PAST sont aussi utilisables pour diverses recherches de biologie et de géologie.

Pourquoi PAST ?

L'introduction au manuel de PAST (http://folk.uio.no/ohammer/past/intro.html) indique plusieurs qualités :

- PAST est gratuit
- PAST est fait sur mesure pour la paléontologie, et diverses fonctions ont été développées spécialement dans ce but.
- PAST est facile à utiliser, et peut servir d'introduction pédagogique à l'emploi de logiciels plus complexes
- le site de PAST montre divers exemples d'utilisation.

On peut ajouter d'autres qualités, qui justifient son emploi pour divers domaines des sciences d'observation :

- PAST fonctionne sous toutes les versions de Windows à partir de Windows 95. On peut aussi l'utiliser sous Linux par l'intermédiaire de WINE (par exemple avec Linux Poseidon).
- PAST est un logiciel léger. Sous forme compactée (zip), il peut tenir sur une seule disquette (moins d'un mégaoctet).
- PAST n'est pas sectaire pour le séparateur décimal. Il peut utiliser aussi bien le point que la virgule, et donc recevoir des données des diverses versions des tableurs.
- PAST est facile à utiliser en complément des tableurs de type OpenOffice Calc ou Excel. Ses fichiers sont des « fichiers-textes » lisibles par les tableurs. On peut aussi copier et coller les données entre le tableur et PAST, de façon très facile.
- En un seul logiciel, PAST regroupe des fonctions de tests statistiques classiques et de visualisation de données qu'on ne trouve en général que dans des logiciels spécialisés.

A qui s'adresse cet ouvrage ?

A tous les scientifiques, professionnels ou amateurs.

- Les spécialistes ont en général des logiciels spécialisés, peut-être plus puissants que PAST. Mais lorsqu'ils sortent de leur spécialisation, et doivent interpréter des données dans un domaine qui n'est plus tout à fait le leur, PAST apporte diverses fonctions permettant d'explorer ces données rapidement.
- Les étudiants en sciences ont souvent des cours de statistiques, mais qui ont un formalisme très mathématique. Ils peuvent avoir besoin de visualiser rapidement des données expérimentales, de faire des comparaison, des graphiques, des estimations... sans trop s'embrouiller dans ces mathématiques.

Structure de cet ouvrage

Globalement, la succession des chapitres correspond à l'ordre des menus de PAST, de gauche à droite. La longueur des chapitres diminue au fil de l'ouvrage, notamment parce que les menus vers la droite sont les plus spécialisés dans des domaines particuliers de la paléontologie, et ont donc moins été développés ici.

Le premier chapitre « Fournir des données à PAST, et les communiquer à d'autres logiciels » correspond aux menus « File » (ouvrir un fichier de données, sauvegarder les données...), au menu « Edit » (copier, couper, coller, ajouter ou retrancher des lignes et des colonnes, transposer un bloc de cellules, etc), et au menu « Transform » (transformation des données par une formule, par tri, par interpolation...).

Le second chapitre « Tracer des graphiques » correspond au menu « Plot ». C'est un ensemble de fonctions très importantes de PAST, avec divers types de graphiques.

Le troisième chapitre « Effectuer des calculs statistiques » correspond au menu « Statistics ». C'est le lieu des estimations de moyennes et de variance, et des tests statistiques plus ou moins classiques.

Le quatrième chapitre « Statistiques multivariables » (menu Multivar) permet de faire analyses en composantes principales, analyses des correspondances, analyses canoniques, mais aussi des classifications en arbres (dendrogrammes) et par regroupement automatique.

Le cinquième chapitre « Modélisation » (menu Model) permet bien sûr la régression linéaire, mais aussi sinusoïdale, polynomiale, logistique...

Le sixième chapitre « Étude de la biodiversité » (menu Diversity) est plutôt destiné aux études d'écologie.

Le septième chapitre « Étude des séries temporelles » est d'un emploi plus large. A partir de séries de données, les fonctions d'analyse spectrale ou d'autocorrélation permettent de déterminer la fréquence des phénomènes périodiques.

Le huitième chapitre « Mesures géométriques » contient des fonctions d'usage assez général, comme les graphiques montrant l'orientation ou la direction de mesures, ou bien l'interpolation pour réaliser des cartes, mais aussi des fonctions plus spécialisées pour la paléontologie comme l'allométrie.

Le neuvième chapitre « Analyses stratigraphiques spécifiques » (menu Strat) est vraiment spécialisé pour la paléontologie et la stratigraphie.

Le dixième et dernier chapitre « Programmation des actions de PaST » (menu Script) indique brièvement comment automatiser des actions de PAST par un langage de programmation.

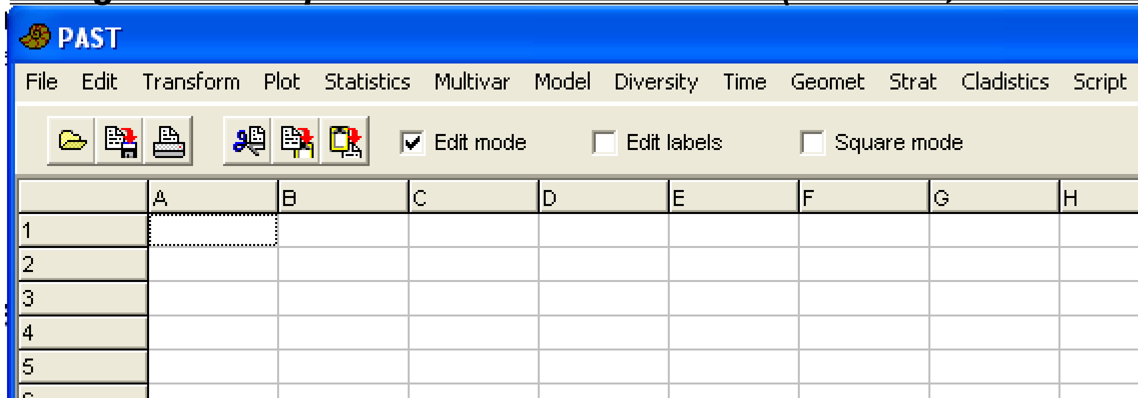
Table des matières

1 Fournir des données à PAST, et les communiquer à d'autres logiciels

PAST ressemble à un tableur : les données sont organisées en lignes (horizontales) et en colonnes (verticales), mais ce n'est pas vraiment un tableur comparable à Excel, OpenOffice Calc ou Gnumeric. Il n'y a qu'une seule table possible, et on ne peut pas mettre de formules dans les cases. On peut certes changer les valeurs par une formule, mais de façon différente des tableurs (voir 1.5).

Les cases ne doivent pas contenir des formules mais seulement des valeurs (numériques ou alphanumériques).

1.1 Le tableau de PAST : les colonnes correspondent aux variables, et les lignes correspondent aux observations (mesures, individus...)



Les cases blanches sont celles qui vont correspondre aux données. Comme dans la plupart des tableurs, les colonnes sont désignées par des lettres, et les lignes sont désignées par des nombres. Au démarrage, il n'y a que 26 colonnes (de A à Z), et 99 lignes, de 1 à 99. On peut augmenter ce nombre par le menu Edit « Insert more rows » pour insérer des lignes et « Insert more columns » pour insérer des colonnes. Lorsqu'on charge un gros fichier, le nombre de lignes et/ou de colonnes est augmenté automatiquement.

Les cases grises correspondent aux labels des lignes et des colonnes. Lorsque « Edit mode »est coché, mais non « Edit labels », on ne peut pas changer le contenu de ces cases. Lorsque « Edit mode » est coché, on peut changer la valeur de ces cases, et ainsi donner un nom explicite aux variables en changeant A, B, C, etc, ou donner un nom explicite aux individus mesurés en changeant 1, 2, 3, etc.

1.2 Cocher ou non les cases « Edition »

1.2.1 Trois cases à cocher au dessous du menu général

– « Edit mode » est coché par défaut. Cela signifie que l'on travaille en mode « édition », et l'on peut modifier le contenu des cases blanches (valeurs mesurées), mais non le contenu des cases grises (labels des lignes et des colonnes). Lorsque cette case n'est pas cochée, on ne peut pas cliquer sur une case et y entrer des valeurs en frappant au clavier, mais on peut quand même coller des données à partir du presse-papier de Windows.

– « Edit labels » n'est pas coché par défaut. Cela signifie que l'on peut changer au clavier la valeur des cases blanches, mais non la valeur des labels de lignes et de colonnes. Lorsque cette case est cochée, on peut aussi changer la valeur des labels de lignes et de colonnes,

aussi bien au clavier que par collage à partir du presse-papier.

- (« Square mode » ne sert qu'à visualiser les cases vides dans certains cas. Le couple présence/absence est codé par 1/0 (respectivement). Toutes les valeurs positives sont considérées comme « présence ». En cochant « square mode », les cases avec présence apparaissent en noir.)

1.2.2 A quoi sert le mode autre qu'édition ?

Lorsque la case « Edit mode » n'est pas cochée, on ne peut pas entrer des valeurs au clavier. Par contre, il est très facile de déplacer des lignes et des colonnes. Il suffit de sélectionner la ligne en cliquant sur son nom (case de gauche), et l'on peut la déplacer simplement en déplaçant le curseur-souris tout en laissant le bouton enfoncé. Il en est de même pour les colonnes, que l'on peut très facilement permuter de la même façon : cliquer sur le nom, puis tirer la colonne d'un côté ou de l'autre en bougeant la souris, bouton gauche enfoncé.

C'est intéressant pour diverses représentations graphiques ou tests statistiques, qui nécessitent d'avoir les colonnes dans un ordre bien déterminé.

1.2.3 Données possibles dans les labels de lignes ou de colonne

Il ne faut pas mettre d'espace dans les étiquettes des colonnes et des lignes (remplacer les espaces par des tirets de soulignement).

1.2.4 Les données possibles dans les cases de mesure :

N'importe quelles valeurs peuvent être entrées dans les cases, mais la plupart des fonctions demandent des valeurs numériques. Le séparateur décimal peut être aussi bien la virgule que le point.

Les données manquantes sont codées par le point d'interrogation « ? » ou la valeur -1. Attention, de nombreuses fonctions ne supportent pas les données manquantes.

Les cellules vides sont codées par un point.

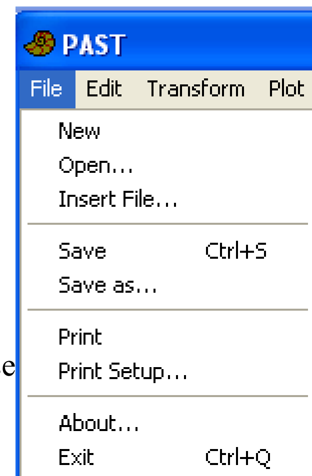
1.2.5 Sélectionner une zone

Attention, cela ne fonctionne pas comme la majorité des tableurs ! PAST se lance normalement en mode « édition » (la case On ne peut sélectionner une zone en déplaçant la souris avec le bouton enfoncé que lorsqu'on n'est pas en mode édition (décocher la case correspondante)

- on sélectionne une ligne en cliquant sur l'étiquette de la ligne (colonne de gauche)
- on sélectionne une colonne en cliquant sur l'étiquette de la colonne (ligne du haut)
- On peut sélectionner plusieurs lignes en cliquant sur l'étiquette de la ligne tout en maintenant la touche majuscule enfoncée.
- On peut sélectionner plusieurs colonnes en cliquant sur l'étiquette de la colonne tout en maintenant la touche majuscule enfoncée.
- On peut sélectionner un pavé de cases en cliquant sur une case, puis sur une autre, tout en maintenant la touche majuscule enfoncée. On peut aussi cliquer sur une case d'un angle du pavé, puis sélectionner ce pavé en appuyant sur la touche « Majuscule » en même temps que sur les touches-flèches.
- On peut sélectionner tout le tableau en cliquant sur la case grise en haut à gauche.

1.2.6 Grouper et colorer des lignes

On a le choix de 16 couleurs et types de puces après activation de l'option Edit | Row color symbol. Cette option permet de visualiser facilement divers groupes de mesures.



1.3 Menu « File » : charger et sauvegarder des données dans des fichiers

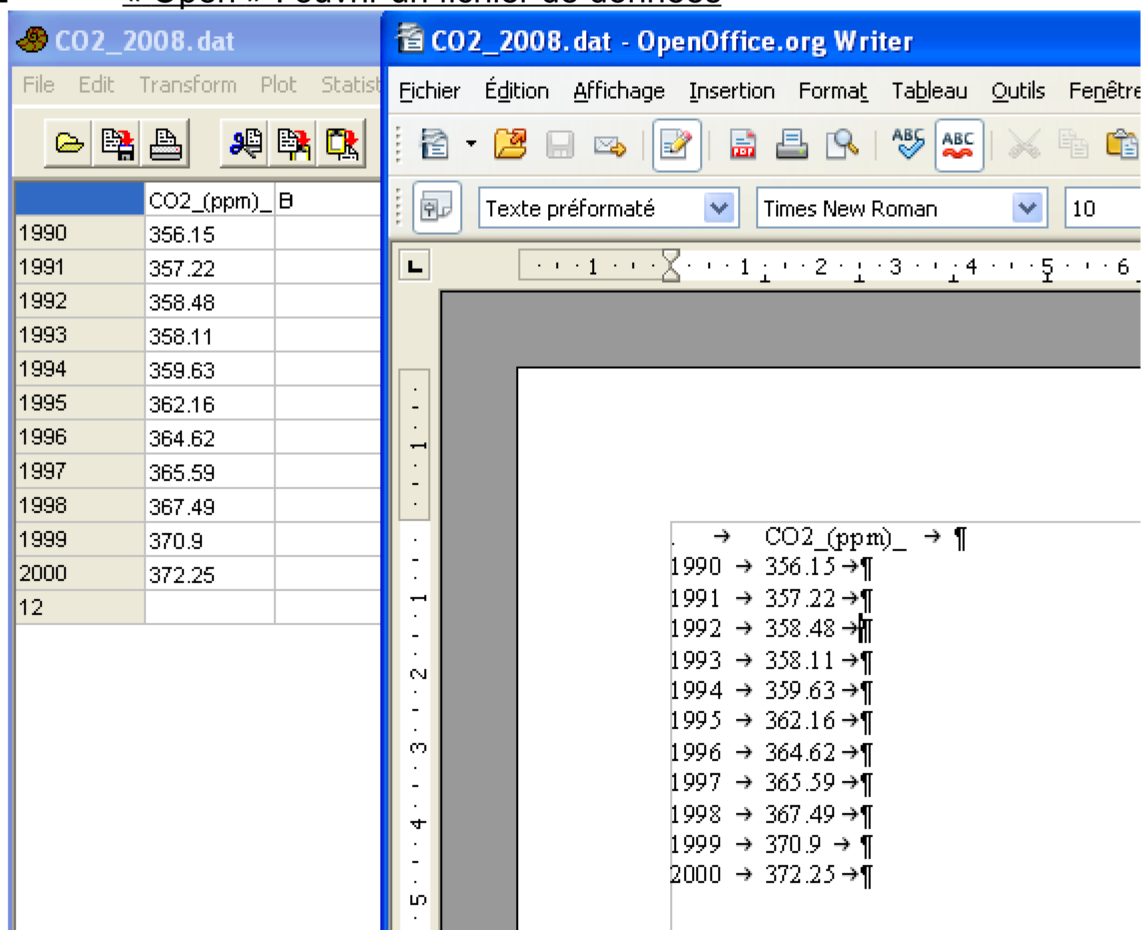
Les fichiers sont de type « texte avec champs séparés par des tabulations ». La fin de ligne (passage à l'enregistrement suivant) est marquée par un retour-chariot et un saut de ligne, ce qui est la norme sous Windows.

Les cellules vides sont codées par des points. Normalement, la première case de la première ligne contient un point.

Les cases ne doivent pas contenir d'espaces, et on doit les remplacer par des tirets de soulignement (ce remplacement est fait automatiquement lorsque PAST lit un fichier-texte provenant d'un autre logiciel (un tableur par exemple), où les espaces sont possibles.

1.3.1 « New » : nouveau tableau vide

1.3.2 « Open » : ouvrir un fichier de données



On peut utiliser la fonction « open » du menu File, ou bien faire un glisser-déposer de l'icône du fichier sur la fenêtre de PAST.

Normalement, PAST peut charger la première page des fichiers Excel ainsi que quelques autres

formats plus exotiques (Nexus, TPS, Rasc...)

MAIS en fait, il faut que le vrai logiciel Excel soit installé sur l'ordinateur : PAST est incapable de charger seul les fichiers Excel. Le mieux est donc d'utiliser les fichiers de type texte, ou bien de coller les données à partir d'un tableur (OpenOffice, Excel, ou autre).

Cette option emplit seulement les premières cases : colonnes à partir de A, lignes à partir de 1.

1.3.3 « Insert File » : pour insérer un fichier dans le tableau actuel de Past

Cette option permet d'ajouter des données dans une partie ou une autre du tableau. Les données sont ajoutées à partir de la case active (vers les colonnes de droite et les lignes au dessous de la case active).

Attention ! les labels des colonnes et des lignes sont bien sûr mis à leur place, en tête de colonne et de ligne, même si de nombreuses cases blanches séparent ensuite les labels des cases contenant les valeurs insérées.

contingence.dat

File Edit Transform Plot Statistics Multivar Model Diversity Time Geomet Strat Cladist

Edit mode Edit labels Square mode

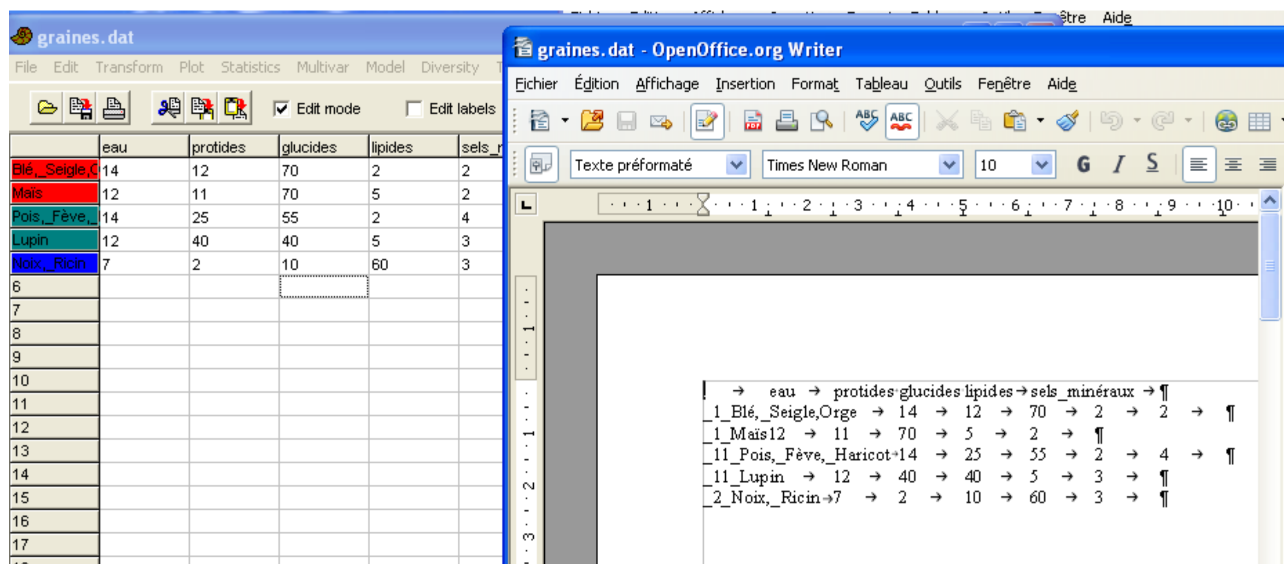
	A	B	C	blanc	bleu	rouge	vert
1							
2							
3							
clair				10	20	30	15
moyen				15	55	44	33
foncé				10	40	35	29
très_foncé				5	20	25	20
8							

1.3.4 « Save » pour enregistrer le fichier

Là encore, le mieux est d'utiliser le format par défaut (.DAT = fichier-texte lisible par les tableurs). Il faut mettre manuellement l'extension « .dat » pour que celle ci existe et soit identifiable ensuite par PAST (fonctions de lecture « Insert file » ou « Open »).

Si des lignes ont été associées à des couleurs, l'étiquette de la ligne dans le fichier commencera par un soulignement, continuera par un nombre de 0 à 15 correspondant à la couleur, et continuera par un soulignement.

Si des colonnes sont d'un type particulier, l'étiquette de la colonne commencera par un soulignement, puis un nombre entre 0 et 3, puis un soulignement (0 : indéfini, 1 = ordinal, 2 = nominal 3 = binaire).



1.3.5 « Save as » pour enregistrer avec un nouveau nom

Là encore, mettre l'extension « .dat » pour que le nom apparaisse dans la liste des fichiers lorsque PAST veut ouvrir un fichier.

1.3.6 « Print » pour imprimer le résultat

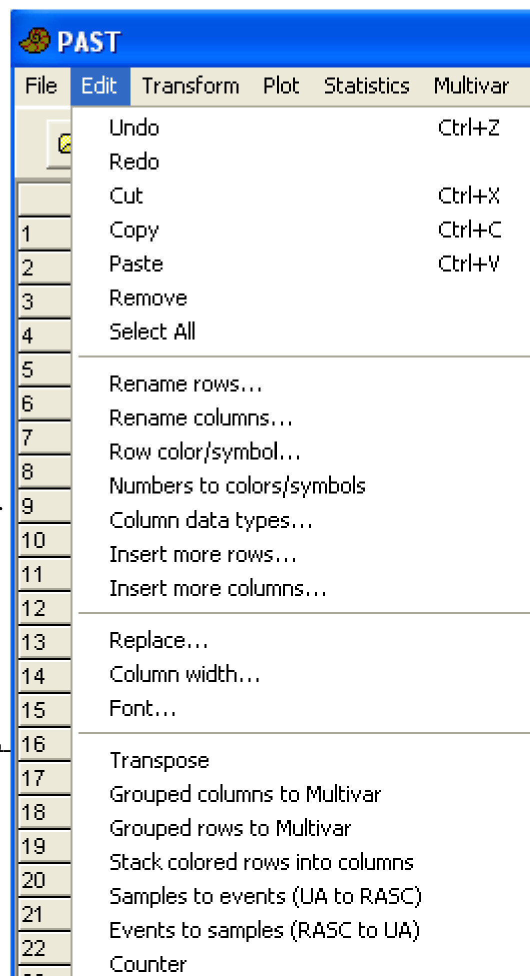
C'est une impression immédiate, sans fioritures, sans réglages de mise en page, sur l'imprimante par défaut du système informatique. Pour une mise en page soignée, utiliser plutôt un bon logiciel de bureautique.

1.3.7 « Print setup » pour régler l'imprimante

Il faut penser à ce réglage avant de lancer l'impression par « Print »

1.3.8 « About » qui indique l'origine de ce logiciel

1.3.9 « Exit » pour clore le logiciel lorsque tout est terminé



1.4 Menu « Edit » : modifier les données d'un tableau

1.4.1 Undo : annuler une opération précédente, et Redo pour la refaire

Il faut comprendre « opération précédente » comme toute action sur le clavier, ou action sur la souris, qui modifie le tableau en cours. Par exemple, coller, couper, frapper une touche pour entrer un chiffre, etc.

Donc lorsqu'on a collé malencontreusement de mauvaises valeurs à la place des bonnes, il faut choisir « Undo » pour revenir à l'état antérieur.

Il y a plusieurs (un grand nombre ?) niveaux d'annulation : en validant plusieurs fois cette option « Undo », on peut revenir à des états antérieurs assez lointains.

1.4.2 Cut, copy, paste pour couper, copier, coller les données sélectionnées

C'est le mécanisme classique pour les logiciels de bureautique. « Cut » pour couper les données et les envoyer vers le presse-papier, pour ne laisser à leur place dans le tableau que des cases blanches. « Copy » pour copier les données vers le presse-papier, mais en laissant les valeurs initiales dans les cases.

« Paste » permet de coller les données à partir du presse-papier. S'il n'y avait qu'une seule valeur dans le presse-papier, elle est collée dans la case active. S'il y avait plusieurs valeurs, toutes ces valeurs sont mises dans les cases suivant la case active, vers la droite et vers le bas. Cette option écrase éventuellement les données qui se trouvaient dans les cases : pour annuler cet écrasement, utilisez l'option « Undo ».

Ce mécanisme d'utilisation du presse-papier est la méthode la plus simple et la plus pratique pour échanger des données avec les logiciels extérieurs, en particulier les tableurs.

1.4.3 Remove pour supprimer une ligne ou une colonne

Pour cela, il faut avoir sélectionné une ligne en cliquant sur son nom (colonne de gauche, en gris), ou avoir sélectionné une colonne en cliquant son nom (ligne supérieure, en gris).

Les autres lignes plus vers le bas, ou les autres colonnes plus à droite, se rapprochent de l'origine et prennent la place des lignes ou colonnes supprimées.

1.4.4 Select All, pour tout sélectionner

Cette option sélectionne l'ensemble des cases du tableau ; c'est la même action que cliquer sur la case-origine (A1), en haut à gauche du tableau.

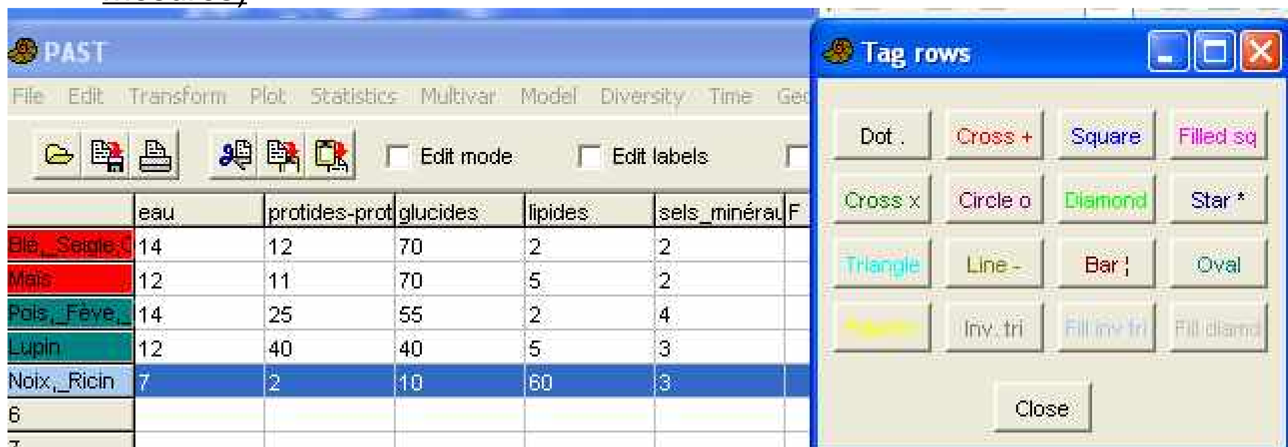
On peut alors couper, copier, coller toutes les données du tableau.

1.4.5 Rename rows et Rename columns pour renommer des lignes ou des colonnes

Lorsque l'on travaille en mode édition (cases « Edit mode » et « Edit labels » cochées), on peut directement cliquer sur les noms des colonnes et des lignes, et les modifier.

Par contre, lorsque la case « Edit labels » est décochée, on ne peut pas modifier ainsi les labels des lignes et des colonnes. Après avoir sélectionné la ligne ou la colonne souhaitée, on valide cette option, et il apparaît une petite boite de dialogue qui permet de modifier ce nom.

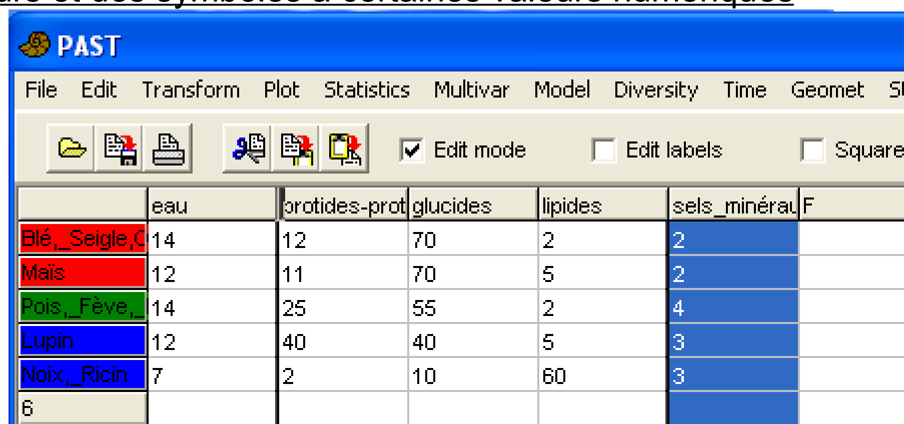
1.4.6 Row color/symbol pour visualiser des groupes d'individus (ou de mesures)



Après avoir sélectionné une ligne (ou quelques lignes adjacentes), cette option permet de colorier la case contenant le nom de la ligne (le label de la ligne). Ceci permet de visualiser des groupes de lignes partageant un caractère en commun. Cette visualisation a 16 modalités possibles, et associe un symbole à une couleur pour chacun des groupes ainsi définis.

Certaines fonctions de visualisation ou de tests statistiques font la différence entre ces groupes de couleurs.

1.4.7 Numbers to colors/symbols : pour associer automatiquement des couleurs et des symboles à certaines valeurs numériques



Après avoir chois une colonne, contenant des valeurs numériques caractéristiques des groupes souhaités, cette option permet d'associer automatiquement une couleur et un symbole à chacun des groupes.

Il faut que la colonne contienne des nombres de 1 à 12. Au delà de 12, la couleur attribuée est équivalente à celle du 1, et pour des nombres négatifs ou nuls, les couleurs sont celles des nombres

entre 1 et 3 (bizarre !).

1.4.8 Column data types : déclarer un type particulier pour les données d'une colonne

Après avoir choisi une ou quelques colonnes, cette option permet de leur attribuer un type particulier de données, qui sera affiché au niveau du label de la colonne :

- ordinal, c'est à dire un classement. La lettre O apparaît dans le label de la colonne
- nominal, c'est à dire une simple description de l'observation. La lettre N apparaît dans le label de la colonne
- binaire, lorsque deux états seulement sont possible : la lettre B apparaît dans le label de la colonne.
- Dans le cas général, le type n'est pas défini, et correspond normalement à des valeurs numériques réelles. Aucune lettre supplémentaire n'apparaît dans le label de la colonne.

1.4.9 Insert more rows et Insert more columns, pour ajouter des lignes ou des colonnes

On peut ainsi frapper au clavier (ou coller à partir du presse-papier) davantage de valeurs, soit pour de nouvelles colonnes, soit pour de nouvelles lignes.

1.4.10 Replace, pour remplacer des caractères ou des valeurs numériques par d'autres

Il apparaît la boite de dialogue classique, permettant de remplacer un caractère par un autre.

PAST accepte aussi bien la virgule que le point comme séparateur décimal. Cette option permet en particulier de remplacer le point par la virgule, ou l'inverse, avant une copie des données à destination d'un autre logiciel plus exigeant pour le séparateur décima, ou bien avant une sauvegarde sous forme de fichier-texte, si on désire le faire relire par un logiciel souhaitant soit le point, soit la virgule.

1.4.11 Column wdth, pour régler la largeur d'une ou quelques colonnes

En dehors du mode « Edit labels », on peut faire varier la largeur des colonnes par la méthode normale des tableurs : cliquer sur la limite droite de la colonne, et faire glisser cette limite tout en appuyant sur le bouton gauche.

Lorsque la case « Edit labels » est cochée, cette méthode n'est pas possible, et il faut employer cette option du menu « Edit », qui fait afficher une boite de dialogue où l'on choisit la largeur de la colonne (en pixels).

1.4.12 Font, pour régler la fonte des caractères d'affichage de l'ensemble de la feuille

Contrairement aux tableurs modernes, il n'est pas possible de choisir la fonte d'une ou quelques cases. Cette option fait apparaître une boite de dialogue où l'on choisit la fonte dans laquelle seront affichées toutes les cases du tableau.

Comme on ne peut pas choisir la hauteur des lignes, cette option est peu intéressante en général.

1.4.13 Transpose, pour transposer l'ensemble du tableau (lignes en colonnes et colonnes en lignes)

En choisissant une fois cette option, l'ensemble du tableau est transposé : les lignes deviennent les

colonnes et inversement. En choisissant une deuxième fois cette option, on revient à la situation initiale, puisque les nouvelles colonnes sont transformées en lignes, ce qui était leur état initial.

Cette option peut être utile pour certaines analyses statistiques, qui donnent des rôles différents aux variables et aux individus.

1.4.14 Grouped columns to Multivar et Grouped rows to Multivar, pour réorganiser les données à partir de fichiers compliqués

Untitled
File Edit Transform Plot Statistics Multivar Model Diversity Time Geomet Strat Cladistics Script
Edit mode Edit labels Square mode

	B	C	D	E
1	0	2.375	4.75	7.109
2	9.484	11.859	14.219	16.594
3	18.969	21.328	23.703	26.063
4	28.438	30.813	33.172	35.547
5	37.922	40.281	42.656	45.031
6	47.391	49.766	52.125	54.5
7	56.875	59.234	61.609	[illegible]
8	66.344	68.719	71.094	73.453
9	75.828	78.188	80.563	82.938
10	85.297	87.672	90.047	92.406
11	94.781	97.156	99.516	101.891

Columns per group
Enter number of columns
4
OK

Untitled
File Edit Transform Plot Statistics Multivar Model Diversity Time Geomet Strat Cladistics Script
Edit mode Edit labels Square mode

	B	C	D	E	E	F
1	0	2.375	4.75	7.109	9.484	11.859
2						
3						

Untitled
File Edit Transform Plot Statistics Multivar Mo
Edit mode

	1
B	0
C	2.375
D	4.75
E	7.109
E	9.484
F	11.859
G	14.219
H	16.594
I	18.969
J	21.328
K	23.703
L	26.063
M	28.438
N	30.813
O	33.172
P	35.547
Q	37.922
R	40.281
S	42.656
T	45.031
U	47.391
V	49.766
W	52.125
X	54.5
Y	56.875
Z	59.234
	61.609

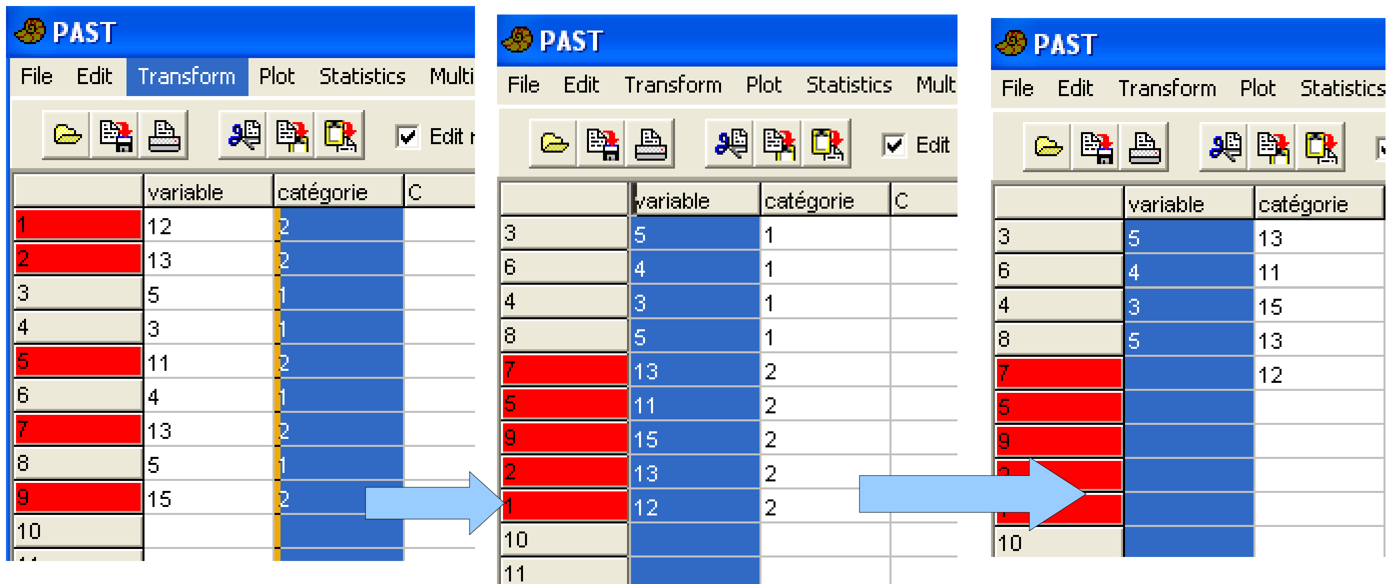
On peut prendre comme exemple les fichiers des logiciels d'acquisition de données Serenis (Jeulin). Pour chaque variable les valeurs sont mises dans un paquet constitué de 4 colonnes ; puis il y a quelques lignes sans valeurs numériques mesurées. De nouveau, une variable dont les valeurs sont en 4 colonnes, quelques lignes, la variable suivante, etc.

Lorsqu'on a copié-collé les valeurs d'une variable dans PAST, et sélectionné les 4 colonnes, on peut prendre l'option « Grouped columns to multivar », et indiquer qu'il y a 4 colonnes. On obtient alors un vecteur-ligne de toutes les valeurs de la variable, que l'on peut transposer par l'option « Edit-Transpose », de façon a obtenir un vecteur colonne de la variable.

En faisant de même pour les diverses variables, puis en collant ces vecteurs-colonnes côte à côte dans un même tableau de PAST, on obtient un tableau constitué des multiples valeurs des diverses variables.

Ce n'est pas simple.... Dans tous les cas, il faut connaître la structure des données.

1.4.15 Stack colored rows into columns : séparer les différentes couleurs en colonnes



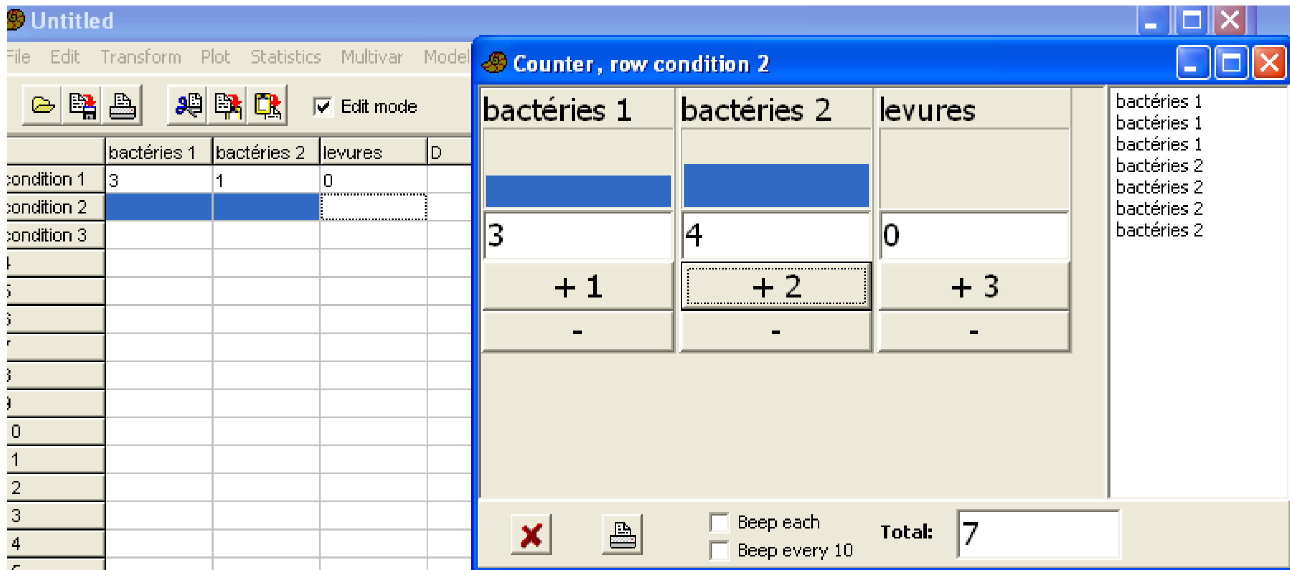
Lorsqu'on a des données de différentes catégories, visualisées par des couleurs différentes, on peut souhaiter étudier les caractéristiques de chaque catégorie. Il faut commencer par trier les données selon ces catégories (menu « Transform | Sort... ». Ensuite, le choix de cette option sépare les catégories en diverses colonnes ; ceci fait, les noms (labels) des colonnes et des lignes n'ont plus de signification, pas plus que les couleurs des lignes.

1.4.16 Samples to events (UA to RASC), et Events to samples (RASC to UA) : à utiliser pour les corrélations biostratigraphiques

Ces deux options sont destinées à être utilisées en biostratigraphie par la méthode des Associations Unitaires (menu « Biostrat »).

1.4.17 Counter : un compteur au clavier

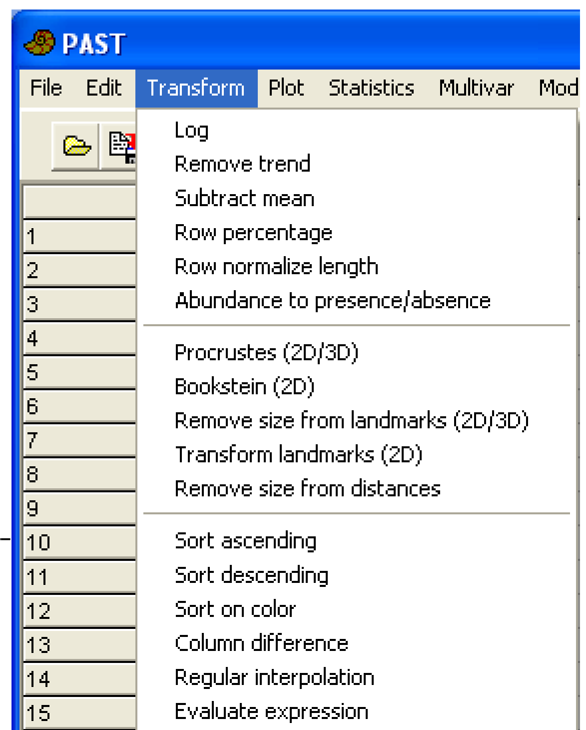
Une fonction de comptage est disponible, par exemple pour compter au microscope des micro-organismes de divers groupes. Chaque groupe va correspondre à une colonne, et une ligne sera utilisée pour chaque étude (tube de Drosophiles, boite de Pétri de colonies bactériennes, quadrat dans une prairie pour le comptage de plantes...)



Par exemple pour le comptage de colonies de deux types différents de bactéries ou de levures, dans diverses conditions :

- sélectionner les trois cases d'une ligne correspondant aux trois types de micro-organismes
- Choisir « Edit |Counter » : une nouvelle fenêtre apparaît.
- En cliquant sur les boutons avec un « + », le compteur de la catégorie correspondante est incrémenté ; en cliquant sur les boutons « - », le compteur est décrémenté.
- Lorsque le comptage est terminé, il faut cliquer sur le bouton marqué d'une croix rouge (en bas à gauche) : la fenêtre se ferme, et les nombres correspondant au comptage apparaissent dans les cases correspondantes du tableau.
- Pour la condition suivante (boite de Pétri suivante, quadrat suivant...), il faut sélectionner dans la ligne suivante les cases correspondant aux types à compter, puis sélectionner « Edit | Counter », etc.

1.5 Transform : Transformer les données



On peut faire subir aux données différentes opérations. Ces opérations ne sont possibles que sur une zone sélectionnée : elles ne sont pas possibles sur une seule case.

1.5.1 Log : logarithme en base 10

Il faut des données positives, mais les données manquantes (?) sont acceptées

1.5.2 Remove trend : soustraction de la tendance (par régression linéaire)

Il faut sélectionner des cases appartenant à au moins deux lignes (éventuellement une colonne ou deux).

Si des cases de deux colonnes sont sélectionnées (X et Y), PAST effectue le calcul de la régression linéaire de Y en X, puis enlève de la colonne Y la valeur de la régression, pour ne plus y laisser que l'écart à la droite de régression. La colonne X n'est pas changée.

Si des cases d'une seule colonne sont sélectionnées, PAST effectue le calcul de la régression linéaire de cette colonne par rapport au numéro des lignes, puis l'enlève de la valeur de chaque case.

Les cases non sélectionnées ne changent pas

1.5.3 Substract mean : soustraction de la moyenne des cases de la colonne

Après avoir sélectionné plusieurs cases (une ou plusieurs colonnes), cette option soustrait à la valeur de chaque case la moyenne des valeurs de la colonne (en fait, la moyenne des cases sélectionnées de la colonne). Les cases non sélectionnées ne changent pas.

1.5.4 Row percentage : pourcentage de la ligne

Après avoir sélectionné plusieurs colonnes, sur une ou plusieurs lignes, cette option remplace les valeurs des cases par le pourcentage de ces valeurs par rapport au total sélectionné sur la ligne. Cela permet de remplacer des quantités absolues par des quantités relatives. Les cases non sélectionnées ne changent pas.

1.5.5 Row normalize length : longueur normalisée sur la ligne

Toutes les valeurs sélectionnées de la ligne sont divisées par la longueur euclidienne de ce vecteur-ligne (la racine carrée de la somme des carrés des valeurs)

1.5.6 Abundance to presence/absence : convertir les données d'abondance en présence/absence

Toutes les valeurs positives sont remplacées par des 1.

1.5.7 Procrustes (2D/3D), Bookstein (2D), Normalize size (2D/3D), Transform landmarks (2D), Burnaby size removal : diverses opérations destinées aux analyses géométriques

Les analyses géométriques sont utilisées dans le menu « Geomet ».

1.5.8 Sort ascending, Sort descending, Sort on color : tri des colonnes

A partir d'observations (les lignes) en désordre, PAST peut faire le tri des lignes du tableau selon les valeurs prises par une variable (une colonne).

Si une seule ligne est sélectionnée, il apparaît un message d'erreur, mais qui ne bloque pas le logiciel. Si plusieurs lignes sont sélectionnées, le tri ne porte que sur ces lignes : les autres lignes ne changent pas. Si toutes les lignes sont sélectionnées, le tri porte sur l'ensemble du tableau.

Si une seule colonne est sélectionnée, c'est cette colonne qui sert de base au tri, soit par ordre croissant (« Sort ascending »), soit par ordre décroissant (« Sort descending »). Si plusieurs colonnes sont sélectionnées, c'est la première colonne (colonne de gauche) qui sert de base au tri.

1.5.9 Column difference

Il faut sélectionner deux colonnes, et la différence entre elles est mise dans la colonne suivante. Attention ! les données éventuelles de la colonne suivante sont écrasées.

1.5.10 Regular interpolation : interpolation pour remplacer des données manquantes

1.5.11 Evaluate expression : évaluer une expression mathématique

Les expressions ressemblent un peu à celles qu'on peut employer dans les tableurs, mais le fonctionnement est différent.

Après avoir sélectionné une zone, lorsqu'on choisit cette option, une fenêtre s'ouvre. Vers le bas de la fenêtre, on doit entrer la formule de transformation des cases sélectionnées selon la syntaxe indiquée dans le haut de la fenêtre. On peut soit frapper la formule au clavier, soit cliquer sur les noms des colonnes, des fonctions, des opérateurs ou des valeurs.

Lorsqu'on clique sur « Compute », la valeur des cases sélectionnées est remplacée par les valeurs de l'expression calculée. Si on reclique sur « Compute », un nouveau calcul est effectué. Par exemple, si on a introduit ln(x), en cliquant une première fois, on remplace la valeur dans les cases par leur logarithme népérien. Si on clique une deuxième fois, on la remplace par le logarithme du logarithme.

- évaluation d'une expression : opérateurs +, -, *, /, ^, mod,
- fonctions abs, atan, cos, sin, exp, ln, sqrt, sqr, round, trunc.
- Les variables suivantes peuvent être utilisées : x (contenu de la cellule), l (cellule de gauche « left », sinon 0), r (cellule de droite « right »), u (cellule supérieure « up »), d (cellule inférieure « down »), mean (moyenne de la colonne), min (valeur minimale de la colonne), max (valeur maximale de la colonne), n (nombre de cellules dans la colonne), i (indice de la ligne), j (indice de la colonne), random (nombre au hasard uniforme entre 0 et 1), normal (nombre au hasard normal, moyenne 0 et variance 1), integral (somme de la colonne courante), stdev (écart-type de la colonne), sum (somme totale de la colonne courante). Les valeurs manquantes (?) sont acceptées.
- On peut aussi entrer des formules tenant compte des valeurs des autres colonnes. Les noms des colonnes sont indiqués dans la colonne de gauche de la partie supérieure de la fenêtre : ils commencent par c_. Lorsqu'on n'a sélectionné que certaines cases d'une colonne, le résultat peut paraître bizarre, parce que la formule fait correspondre la première ligne de la

zone sélectionnée avec la première ligne de la colonne indiquée dans la formule.

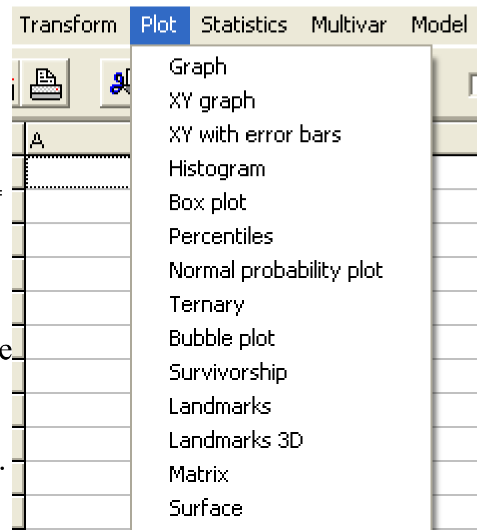
1.6 PAST utilisé sous Linux

PAST est fondamentalement conçu pour être utilisé dans l'environnement MS-Windows, mais l'utilitaire WINE permet de l'utiliser sous Windows. Il fonctionne en particulier très bien sous Poseidon, qui est un « linux scientifique » dérivé de Ubuntu : il suffit de cliquer sur l'icône Past.exe pour que PAST se lance convenablement, puisse bien lire les fichiers, effectuer les calculs et tracer les graphiques.

Les fonctions importantes de PAST fonctionnent correctement, en particulier l'entrée et la sortie de données par lecture et écriture de fichiers, mais aussi par copier/coller, ainsi que la sauvegarde des graphiques dans les formats .emf, .jpg et .bmp.

Par contre, les fonctions qui interagissent trop fortement avec le système d'exploitation ne sont pas disponibles. On ne peut pas copier le graphique dans la mémoire pour le coller directement dans un logiciel de dessin, ni imprimer, ni choisir la fonte de caractères pour les graphiques.

2 Plot : Tracer des graphiques



2.1 Pour tous les graphiques : cliquer pour régler les préférences du graphique

Pour les différents types de graphes : lorsqu'on double-clique sur le graphe, une fenêtre de réglage apparaît, semblable pour tous les graphiques. Certaines des options ne sont valables que pour certains graphiques particuliers, et il ne sert à rien de les modifier pour les autres graphiques.

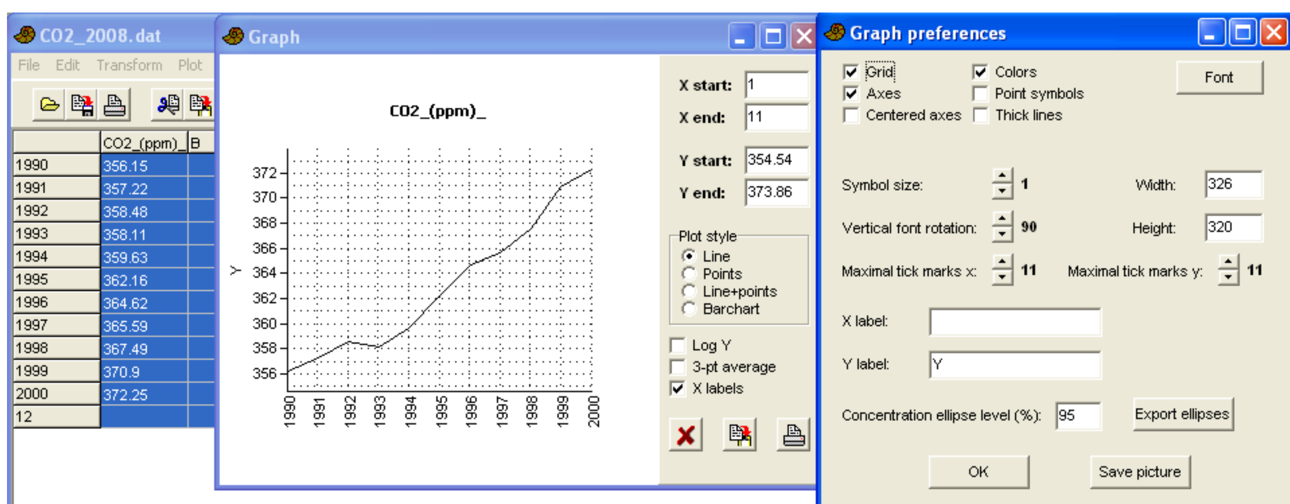
On peut régler l'existence ou non d'une grille ou des axes, l'existence ou non de couleurs (pour les graphiques qui le permettent), l'épaisseur de ligne, la fonte des caractères...

Lorsque le graphique montre des symboles pour les points de mesure, on peut régler la taille de ces symboles, lorsque les axes ont des graduations, on peut régler l'écartement entre ces graduations, mettre des légendes aux axes (« X label » et « Y label »).

On peut sauvegarder le graphique dans les formats emf, jpg et bmp.

Le format emf est un format vectoriel, ce qui permet d'avoir des traits nets même lorsque le graphique est fortement agrandi. On peut aussi le modifier par des logiciels de dessin vectoriel tels que OpenOfficeDraw, mais ce n'est pas très pratique car les différents constituants du dessin ne sont pas groupés.

Les formats jpg et bmp sont des formats bitmaps, et la pixellisation devient visible lorsqu'on agrandit l'image. Pour éviter ceci, il faut agrandir le graphique tracé par PAST avant de le sauvegarder (mais, bien sûr, la taille du fichier augrmente).



Dans la fenêtre du graphique lui-même, en bas à droite, trois icônes permettent de fermer la fenêtre, de copier le graphique dans le presse-papier, et d'imprimer le graphique.

2.2 Graph : simples graphes de position

Conseil : cocher la case X labels, pour que les titres des lignes soient affichés.

Par défaut, le graphique est tracé en simple ligne (« Line »), mais une série de boutons-radio permet d'avoir un aspect différent : « Points » ne trace que les points de mesure, « Line+points » trace à la fois la ligne et les points, « Barchart »devrait tracer un diagramme en bâtons, mais le résultat est illisible.

Cocher « Log Y » fait une transformation en logarithme décimal pour l'axe vertical, et cocher « 3-pt average » fait un lissage de la courbe sur 3 points.

Par défaut, les axes sont optimisés pour que la courbe ou les points occupent le plus de place. On peut régler le début et la fin de l'axe Y par les cases « Y start » et « Y end ». On peut aussi régler le début et la fin de l'axe X par « X start » et « X end », mais cette option n'a guère d'intérêt ici, contrairement aux graphiques suivants.

2.3 XYgraph : le graphique cartésien traditionnel

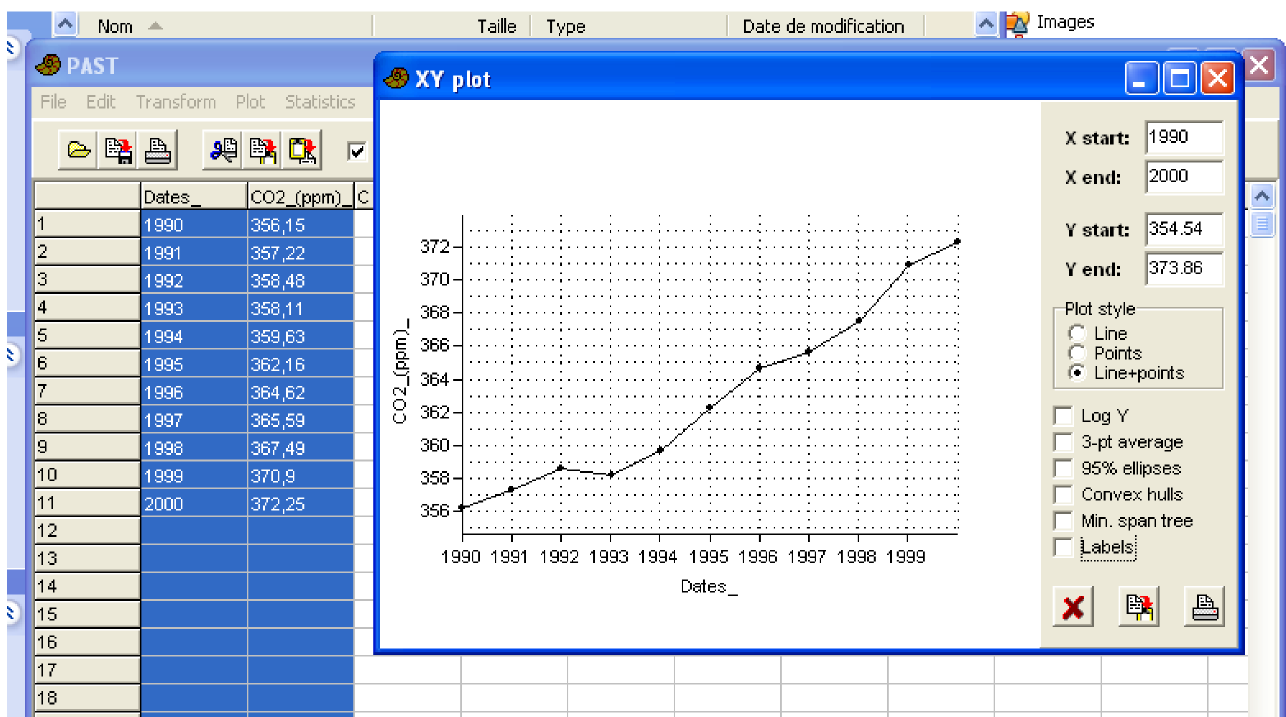
La colonne de gauche donne l'axe des abscisses, la colonne de droite donne l'axe des ordonnées. On peut régler l'étendue des axes par les 4 lignes de saisie en haut à droite de la fenêtre.

Les mêmes réglages que pour le graphique précédent sont possibles, avec en plus :

– « 95% ellipses » qui est l'ellipse de confiance des données au seuil 95% (on suppose que les deux variables ont une distribution normale). Le seuil peut être réglé dans la fenêtre des préférences du graphique.

– « Convex hulls »qui trace l'enveloppe de l'ensemble des points de mesure.

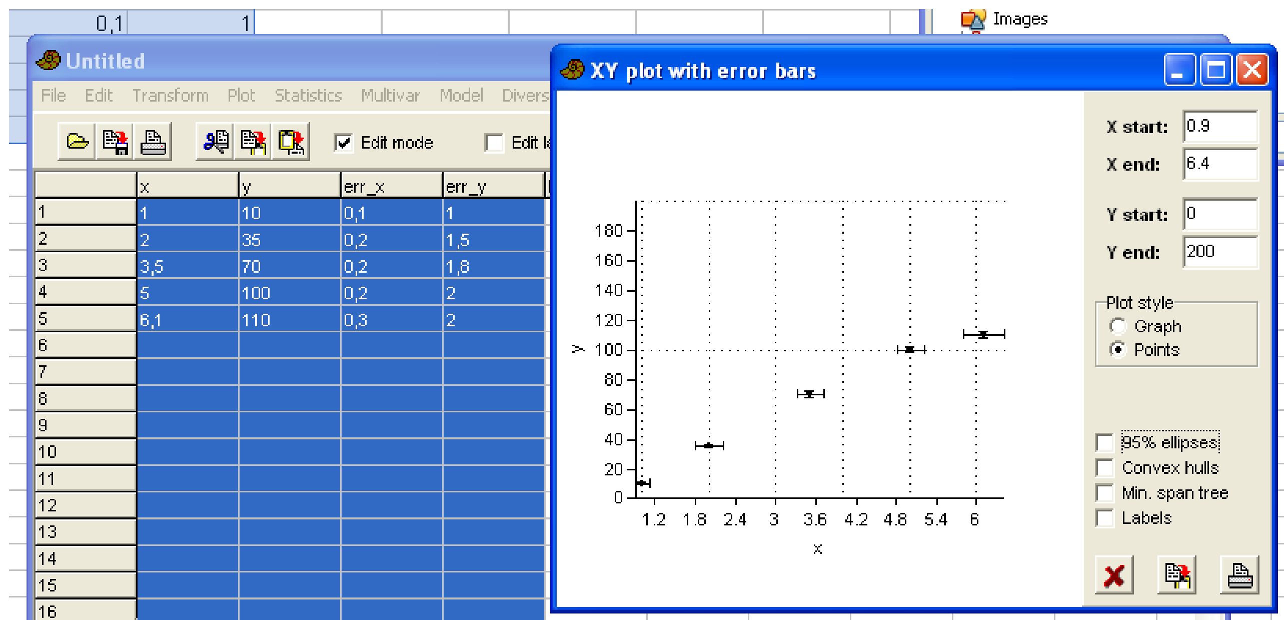
– « Min.span tree » est le chemin minimal joignant l'ensemble des points de mesures.



2.4 XY with error bars : graphique XY avec barres d'erreur

Il faut définir quatre colonnes, de la gauche vers la droite : abscisses X, ordonnées Y, erreur en X,

erreur en Y.

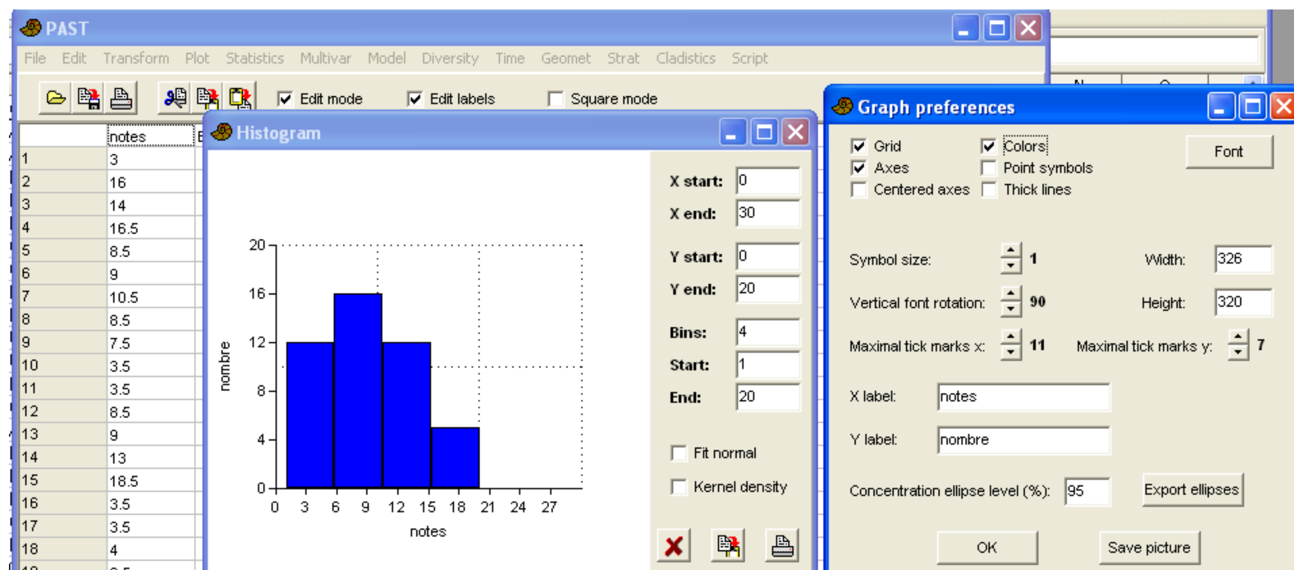


2.5 Histogram : histogramme de fréquence absolue (nombre d'individus) de la population

La case « Bins » indique le nombre d'intervalles (compartiments) désirés, qui était mis par défaut à une valeur « optimale ». On peut aussi changer le début et la fin des compartiments désirés.

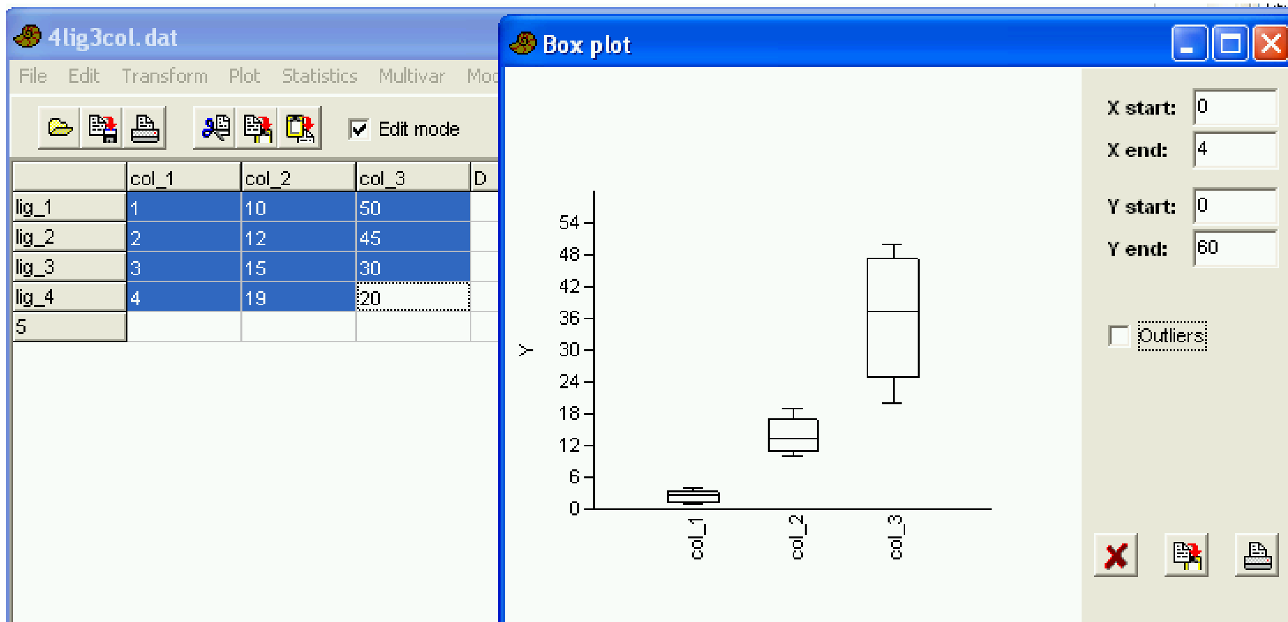
Cocher la case « Fit normal » pour tracer une distribution normale, de paramètres (moyenne et écart-type) identiques à la population dont on vient de tracer l'histogramme.

La case « Kernel density » trace un autre lissage de l'histogramme.



2.6 Box-plot : boites à moustaches = diagrammes de Tukey

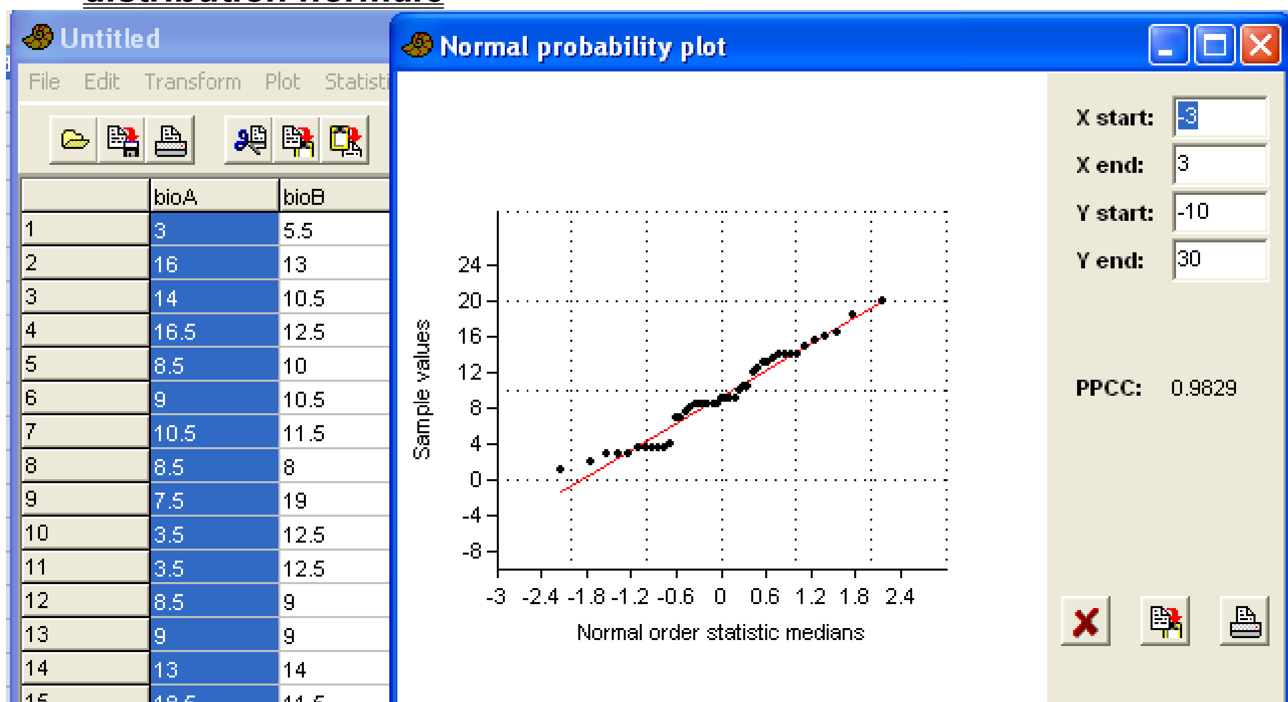
Chaque colonne correspond à une « population », et le graphique présente des boites entre les deux quartiles extrêmes, et des moustaches jusqu'au maximum et minimum.



Cocher la case « Outlier » provoque un autre mode de traçage des moustaches. Les moustaches vont jusqu'au maximum et au minimum, à condition qu'ils ne soient pas trop écartés des quartiles (la moustache ne doit pas faire plus de 1,5 fois la longueur du segment de boite) ; si des points sont en dehors de cet intervalle, ils sont tracés sous la forme de petits cercles, indiquant que ce sont des valeurs exceptionnelles (points aberrants ?).

2.7 Percentiles : diagramme de fréquences cumulées

2.8 Normal probability plot : graphique montrant (ou non) une distribution normale



C'est une sorte de graphique en XY, où l'axe X correspond à une distribution normale et l'axe Y correspond aux valeurs observées. En rouge est une droite de régression entre ces deux

distributions, et à droite « PCCC » est le coefficient de corrélation linéaire entre ces deux distributions.

Ce type de graphique est utilisé pour répondre aux questions du type :

- Les données suivent-elles une distribution normale ?
- Quelle est la nature de l'écart à la normalité : asymétrie, queues plus courtes ou plus longues que normalement ?

Si les données sont distribuées normalement, les points suivent la droite de régression

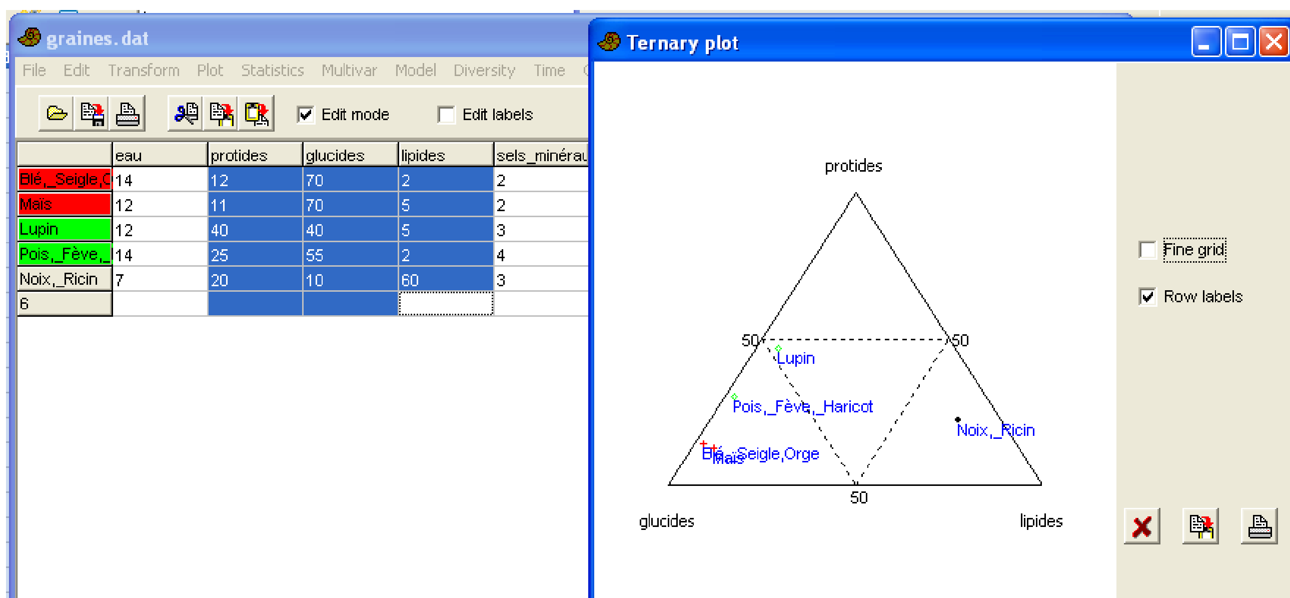
Si la distribution a des queues plus grosses que normalement, les points forment une sigmoïde (à droite ils sont plus bas que la droite, à gauche ils sont plus hauts).

Si la distribution a des queues plus maigres que normalement, les points forment une tilde (à droite ils sont plus haut que la droite, à gauche ils sont plus bas).

Si la distribution est décalée vers la droite, les points forment une courbe au dessous de la droite.

Si la distribution est bimodale, la ligne des points noirs fait deux ondulations à travers la droite rouge...

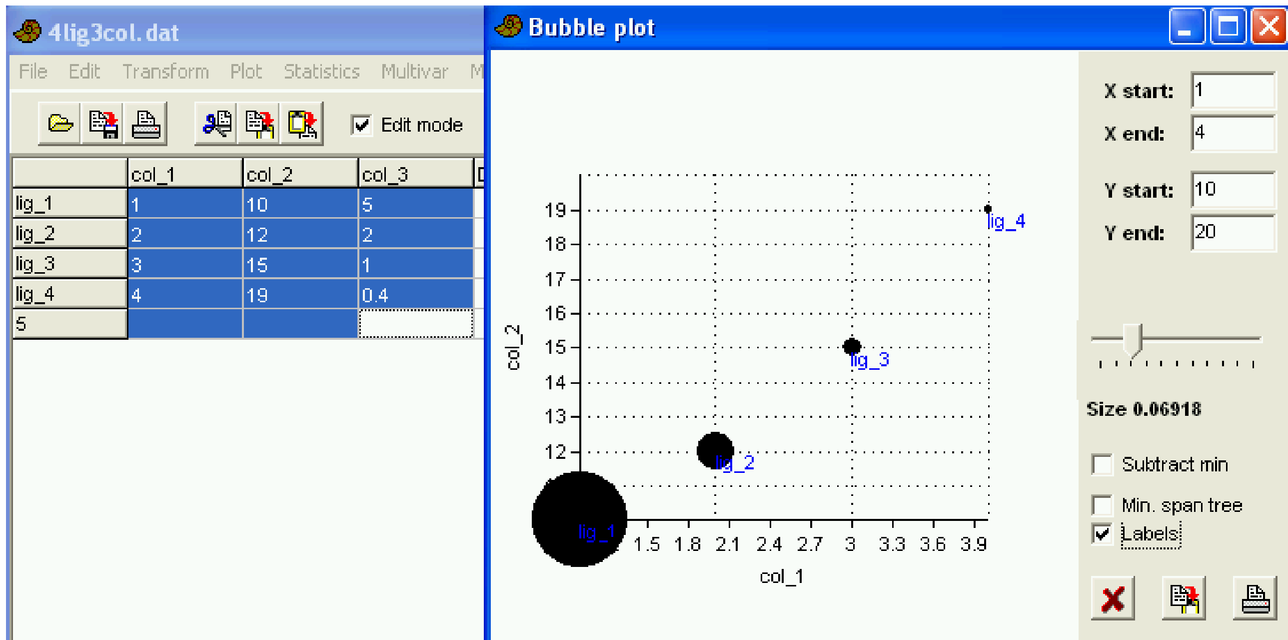
2.9 Ternary : diagramme triangulaire (= diagramme ternaire), pour la composition en 3 facteurs



On peut facilement visualiser les groupes de données en utilisant l'option de coloration des lignes (menu Edit)

2.10 Bubble plot = graphique à bnulles

C'est un simple graphique en XY selon les deux premières colonnes, mais où la taille des points est fonction de la valeur de la troisième colonne. Le curseur « Size » permet de faire varier le diamètre des points pour que l'ensemble reste lisible.



Les valeurs négatives ne sont pas tracées, puisqu'une largeur négative d'un disque n'aurait aucun sens. Cliquer sur la case « Substract min » permet d'affecter une valeur 0 au minimum, et des valeurs positives aux autres valeurs.

2.11 Survivorship : courbes de survie d'une population

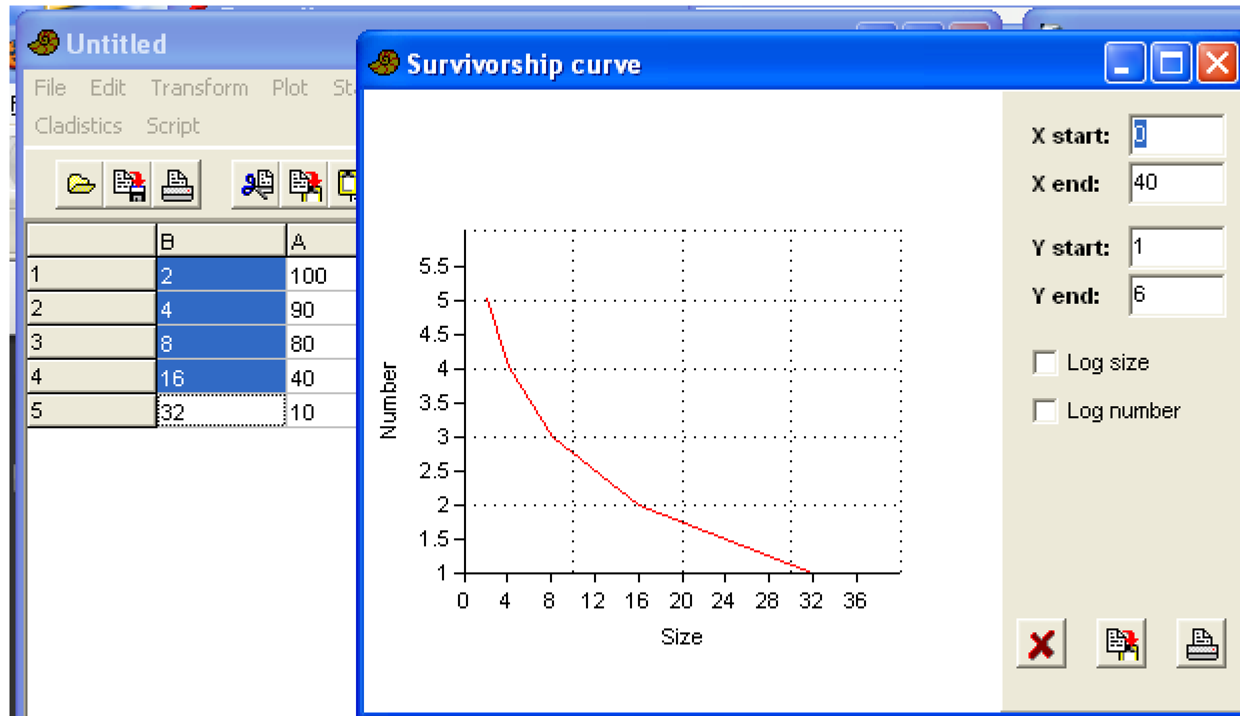
Fondamentalement, la courbe de survie est une courbe figurant la proportion d'individus vivants en fonction du temps (ou d'une dose de traitement).

D'habitude, on trace le nombre d'individus en Y, en fonction du temps en X.

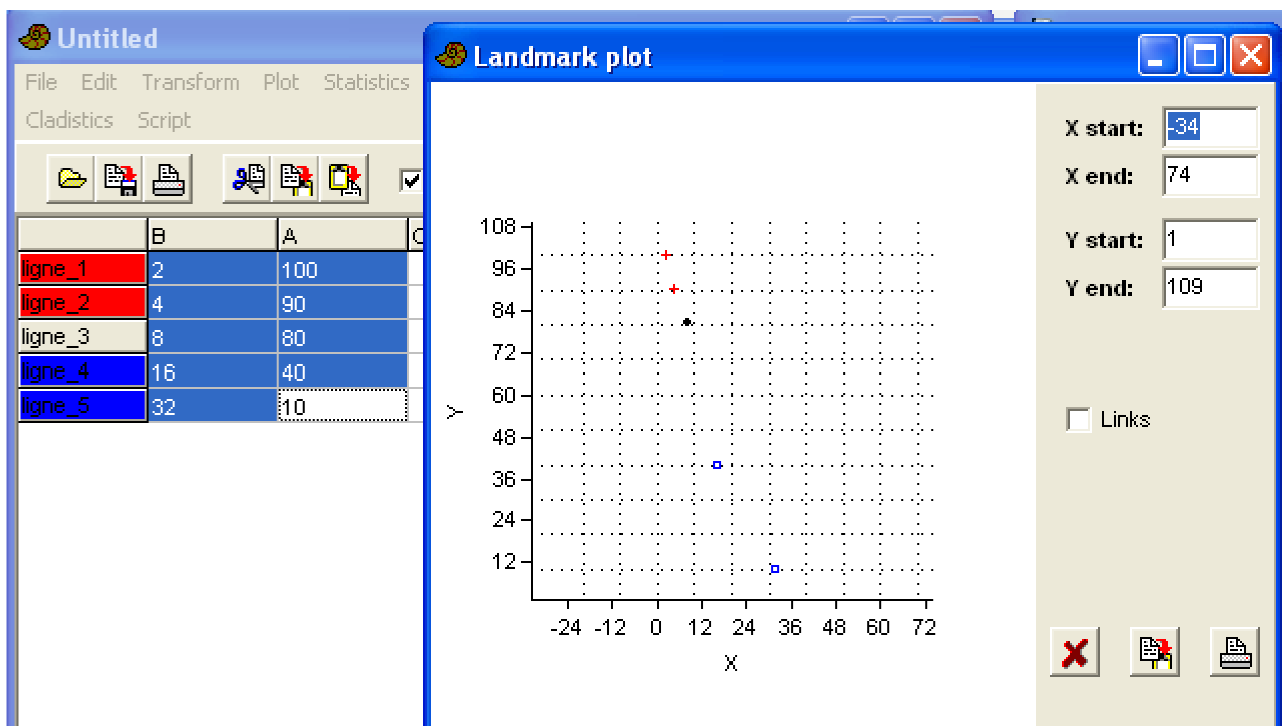
Ici :

- en abscisse est la taille de la population (nombre d'individus)
- en ordonnées est le temps (nombre de générations, ou nombre de jours, ou nombre d'années;..)

Dans tous les cas, les courbes descendent.



2.12 Landmarks : graphique XY avec points de repères par colorations

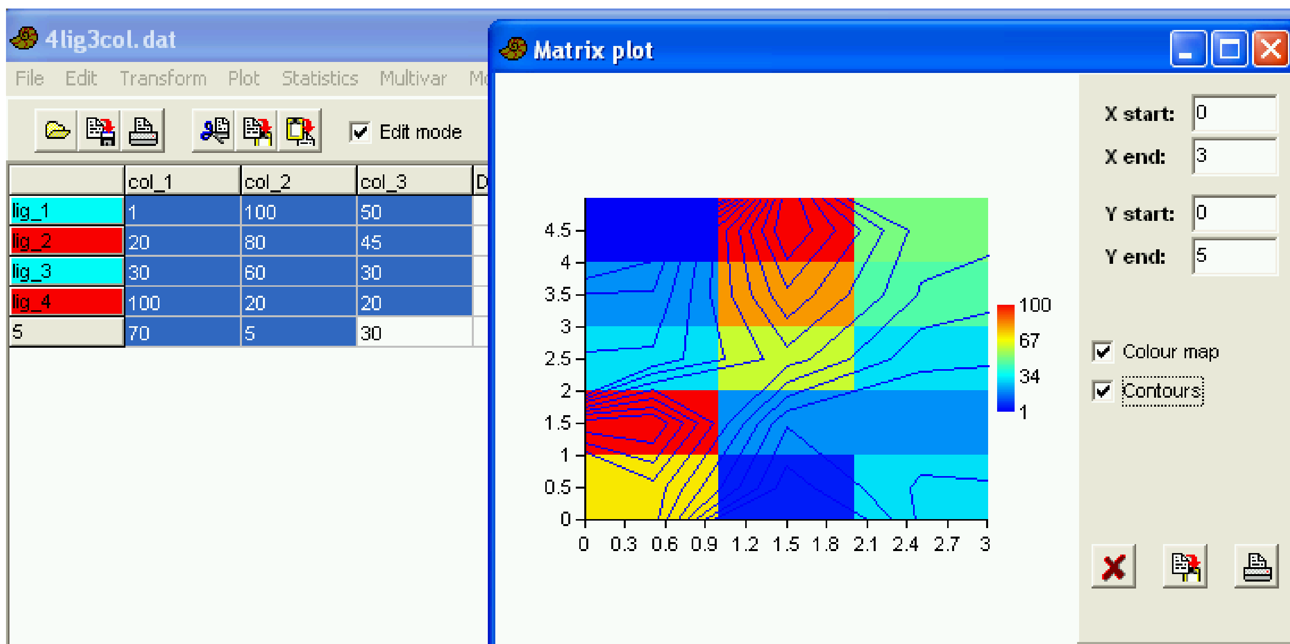


2.13 Landmarks 3D : graphique XY avec une troisième dimension, peu claire

2.14 Matrix : cartographie des valeurs, soit en nuances de gris, soit en couleurs

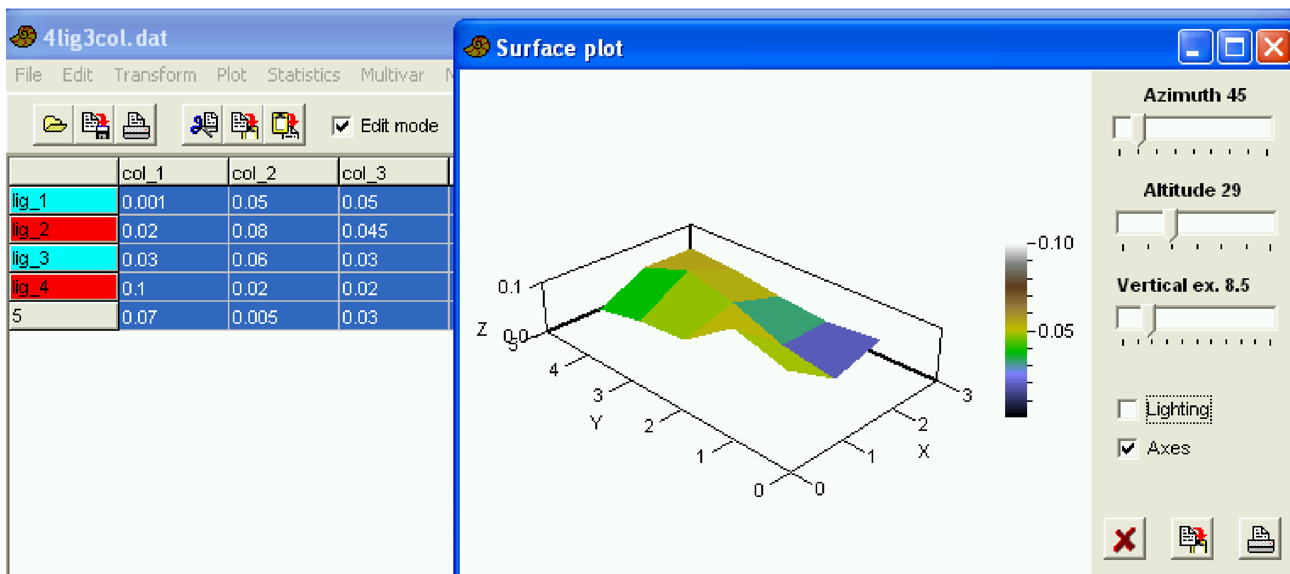
Cette option est intéressante pour figurer des cartographies (pollutions, concentrations, abondance

d'une espèce, etc). L'axe des abscisses est le numéro de la colonne, l'axe des ordonnées correspond aux lignes, et la valeur de la couleur correspond à la valeur donnée dans chaque case. « Contours » permet de tracer des estimations de valeurs intermédiaires.

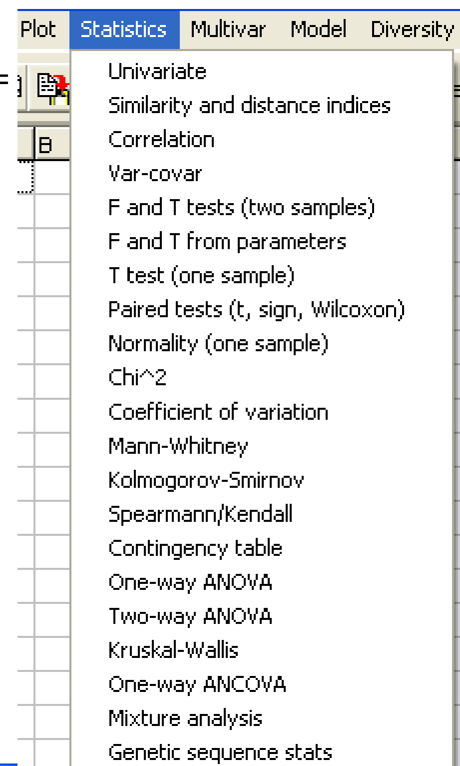


2.15 Surface : cartographie en relief

Là aussi, les axes horizontaux X et Y sont les numéros des lignes et des colonnes, et l'axe Z correspond aux valeurs numériques qui sont dans les cases. Cet axe Z est visualisé d'une part par la forme de la surface (relief), d'autre part par les couleurs.



3 Statistics : effectuer des calculs statistiques



3.1 *Univariate : calculs sur chaque colonne indépendamment*

N désigne le nombre d'observations : soit la totalité de la colonne, soit le nombre de lignes sélectionnées. En cas de données manquantes, elles doivent être indiquées par le point d'interrogation dans les cases.

Min désigne le minimum de la colonne, Max le maximum, Sum la somme des valeurs, Mean la moyenne, Std. error l'erreur standard de la moyenne (écart-type divisé par la racine du nombre d'observations), Variance la variance, Stand. dev l'écart-type, Median la médiane, 25 prcntil et 75 prcntil les premier et troisième quartiles, Skewness et Kurtosis les coefficients

Untitled

File Edit Transform Plot Statistics Multivar Model Diversity Cladistics Script

Edit mode Edit labels

	bioA	bioB	mathA	mathB	physic
1	3	5.5	8	17	9.5
2	16	13	18	13.5	12
3	14	10.5	12	12	18
4	16.5	12.5	13.5	16.5	14
5	8.5	10	19	14.5	16
6	9	10.5	11	16.5	19
7	10.5	11.5	7.5	2.5	7
8	8.5	8	3.5	1.5	7
9	7.5	19	19.5	20	15
10	3.5	12.5	6.5	7.5	8.5
11	3.5	12.5	3.5	8.5	5
12	8.5	9	6.5	17	13.5
13	9	9	4.5	6	8.5
14	13	14	12.5	9	12.5

Univariate statistics

	bioA	bioB	mathA	mathB
N	43	43	43	43
Min	1	4	2.5	1.5
Max	20	19	19.5	20
Sum	402.5	443.5	412	420.5
Mean	9.36047	10.314	9.5814	9.77907
Std. error	0.735947	0.490221	0.705721	0.736507
Variance	23.2896	10.3336	21.4158	23.325
Stand. dev	4.82593	3.21459	4.62772	4.8296
Median	9	10	8	8.5
25 prcntil	4	8	6.5	6
75 prcntil	13.5	12	12.5	14
Skewness	0.155484	0.613793	0.671844	0.428792
Kurtosis	-0.748051	0.77153	-0.353626	-0.701804
Geom. mean	7.83096	9.8224	8.50167	8.49581

d'asymétrie et d'aplatissement par rapport à une distribution normale, Geom. mean la moyenne géométrique.

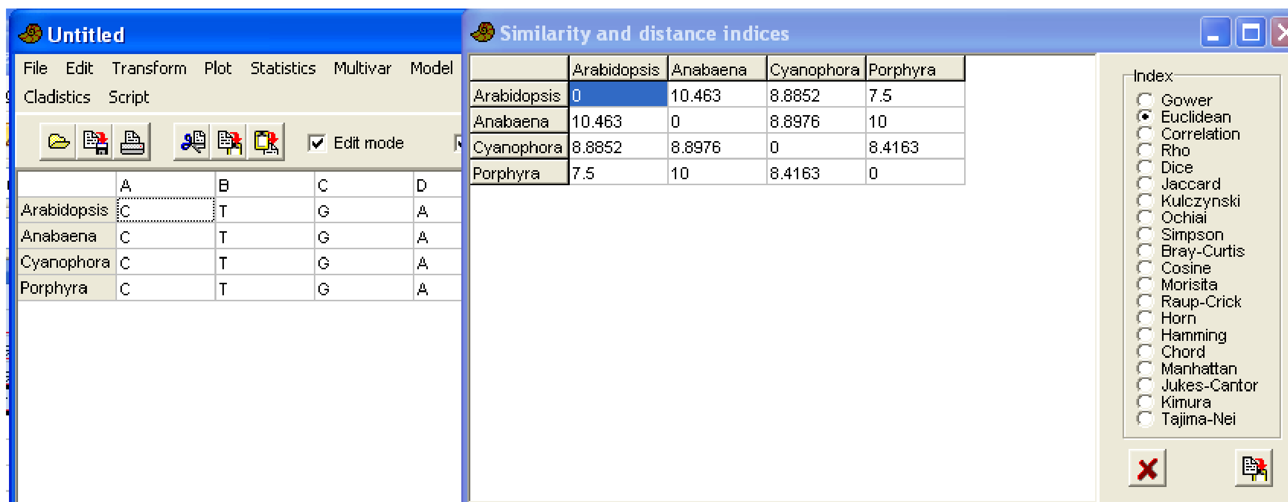
Le coefficient d'asymétrie est positif si la partie droite de la distribution est plus épaisse.

Le coefficient d'aplatissement (qu'il vaudrait mieux nommer coefficient de pointicité ?) correspond au paramètre statistique $\gamma_2 = \mu_4/\sigma^4 - 3$. Il est positif si la courbe de la distribution est pointue. Dans la pratique, on considère que s'il est compris entre -2 et + 2 on peut considérer que la distribution a un aplatissement normal, et que s'il est supérieur à 2, la distribution est plus pointue que normalement, et s'il est inférieur à -2, la distribution est aplatie.

3.1.1 Similarity and distance indices : distances ou corrélations entre les individus

Attention ! D'habitude, on cherche les corrélations entre les variables (les colonnes), à partir d'un certain nombre de mesures (les lignes). Ici, on obtient les corrélations entre les individus, c'est à dire la ressemblance entre les individus pour ce qui est d'un certain nombre de caractères.

Au lieu de valeurs numériques, on peut utiliser des séquences d'acides nucléiques, avec les valeurs A, T, C, G, U (et ? pour les nucléotides absents).



Par défaut, c'est la distance euclidienne qui est calculée, mais il existe une grande variété d'autres calculs possibles, dont le détail est indiqué dans le fichier http://folk.uio.no/ohammer/past/past.pdf.

Ces résultats peuvent servir de base pour la construction de dendrogrammes (« graphiques en arbres de différences ou de ressemblances »), mais il existe d'autres méthodes : voir le menu « Multivar », et en particulier « Multivar | cluster analysis » (4.10).

3.2 Correlation : corrélation entre les variables

Il faut sélectionner plusieurs colonnes d'un tableau où les colonnes correspondent aux variables observées, et les lignes aux individus (lieux, mesures...) où l'on observe ces variables.

Cette fonction fait apparaître une matrice carrée où les cases contiennent les corrélations entre les variables.

3.3 Var-covar : calcul de la matrice de variance-covariance des colonnes

A partir du même tableau que la fonction précédente « Correlation », cette fonction fait apparaître une matrice carrée où les cases diagonales contiennent les variances de chaque variable, et les autres cases contiennent les covariances de chaque couple de variables.

3.4 F and T tests (two samples) : comparaison des variances et des moyennes de 2 échantillons de distributions normales

La structure du tableau est différente des fonctions précédentes. Ici, il n'y a qu'une seule variable par tableau, par exemple la longueur du corps, et deux colonnes contenant les valeurs mesurées. Chaque colonne correspond à un échantillon et les diverses lignes correspondent aux diverses mesures réalisées pour l'échantillon en question. Ce nombre de lignes peut être différent selon les colonnes.

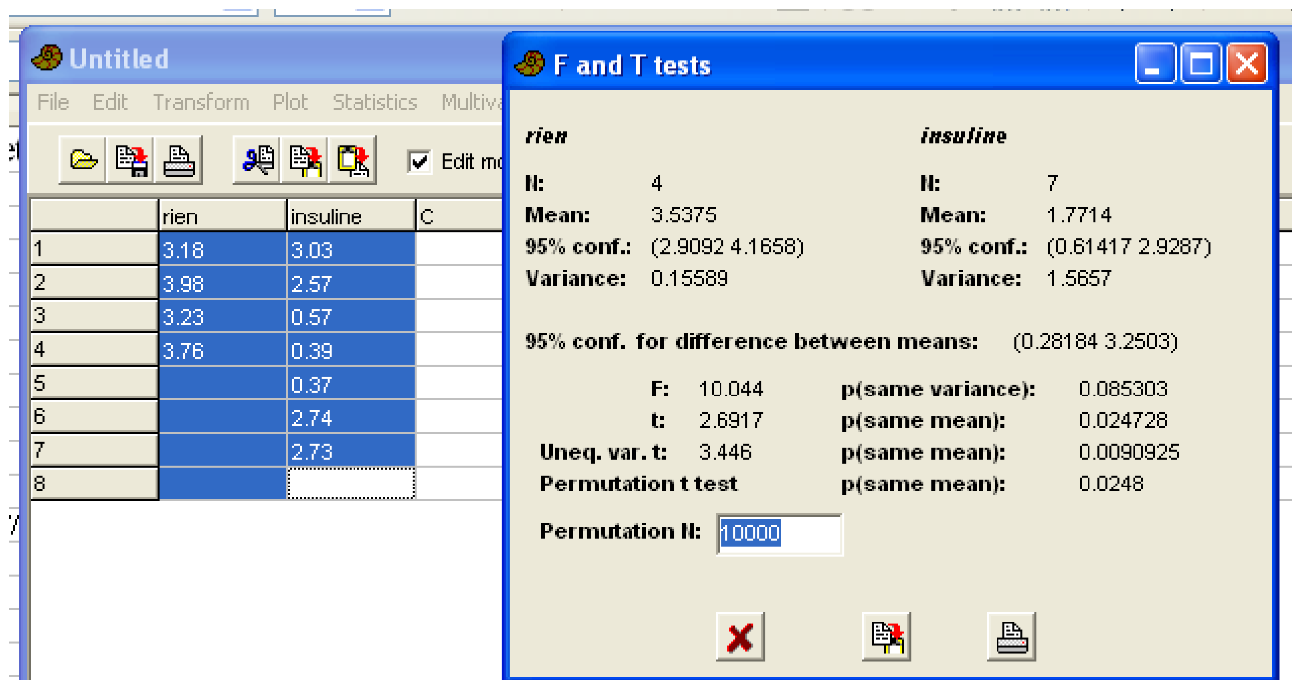
Pour chaque colonne (chaque échantillon), cette option donne :

- N : le nombre d'individus mesurés
- Mean : la moyenne
- « 95% conf. » : l'intervalle de confiance de la moyenne à 95% (plus le nombre de mesures est grand, plus l'intervalle de confiance de la moyenne se rétrécit : l'estimation de la moyenne est plus précise)
- variance : la variance

Le test F fait la comparaison des variances. « p(same variance) » donne la probabilité que les variances observées sur les échantillons soient identiques : si c'est inférieur à 0,05, on peut conclure que les variances sont différentes, et que les deux échantillons proviennent de deux populations différentes (on suppose que les distributions sont normales).

Le test t fait la comparaison des moyennes. « p(same mean) » donne la probabilité que les moyennes soient identiques (simplement sous l'effet du hasard, à partir d'une même population). Si c'est inférieur à 0,05, on peut conclure que les moyennes sont différentes (que les deux échantillons proviennent de deux populations différentes). Ce test suppose que les deux échantillons ont une distribution normale et une variance semblable

Pour des petites populations, ou pour des populations de distribution non normale, il faut mieux utiliser le test « permutation t test).



3.5 *F and T from parameters : comparaison de deux échantillons à partir des paramètres statistiques*

C'est le même principe de comparaison de variance et de moyenne entre deux échantillons, mais dont on ne connaît pas les valeurs mesurées pour chaque individu. Il faut donc entrer à la main les valeurs numériques de la moyenne, de la variance et du nombre d'individus mesurés.

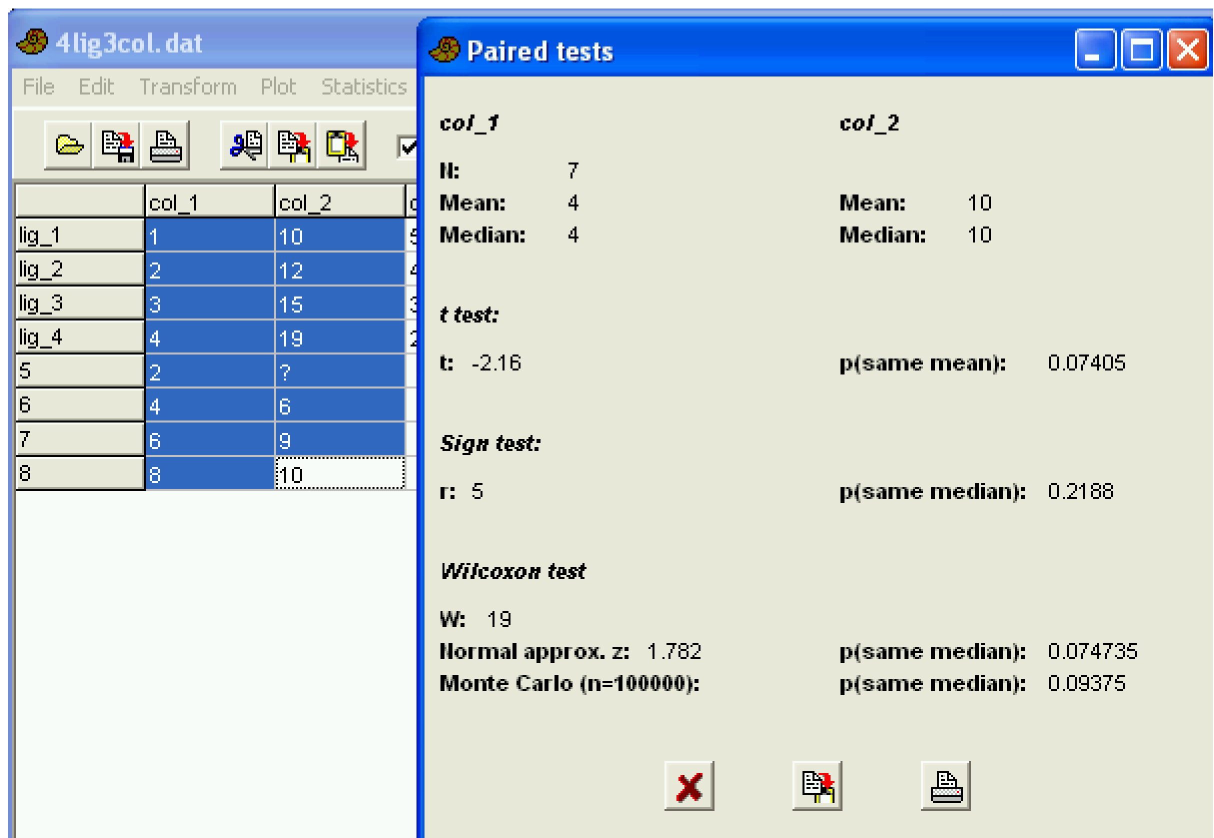
3.6 *T test (one sample) : comparaison d'un échantillon avec une distribution théorique*

On sélectionne une colonne (un échantillon), et cette option permet de comparer à une vaste population dont on connaît la moyenne.

Il faut entrer la moyenne dans la ligne de saisie, et cliquer sur « Compute » pour lancer les calculs.

3.7 *Paired tests (t,sign, Wilcoxon) : comparaison d'échantillons appariés*

« échantillons appariés » signifie qu'à chaque mesure d'une colonne est associée une mesure de l'autre colonne, sur la même ligne : chaque ligne correspond à un individu (un lieu de mesure, une expérience, un village, etc.). C'est par exemple la comparaison d'animaux avant et après engraissement : c'est le même animal avant et après.



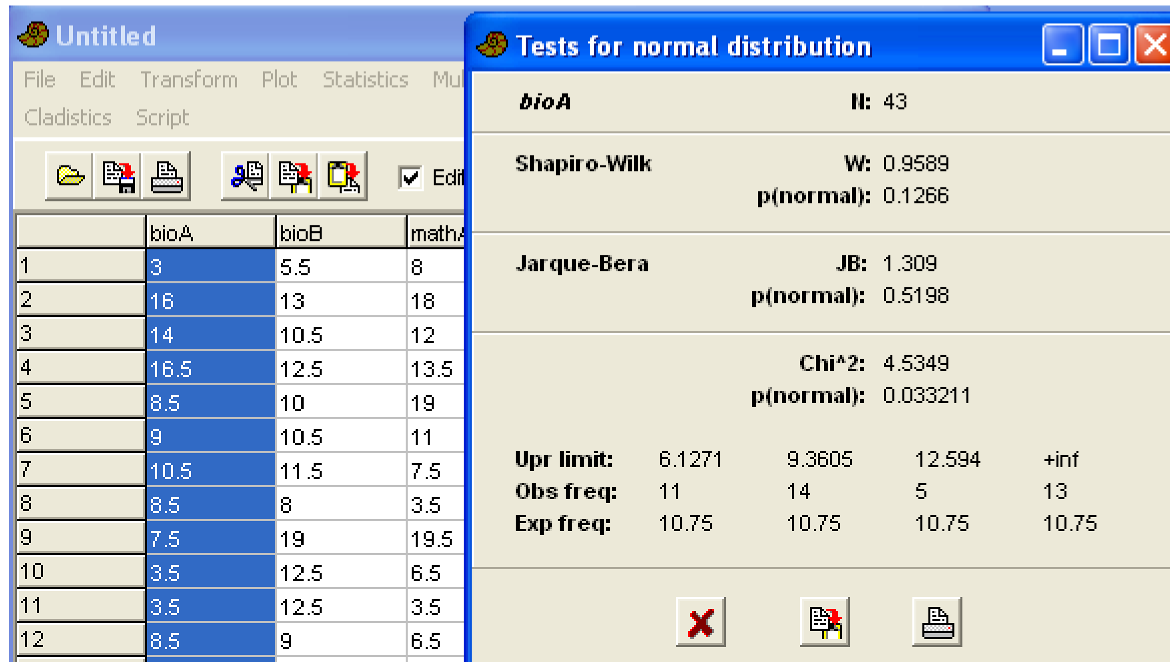
Il faut obligatoirement le même nombre de lignes pour les deux colonnes, mais on peut avoir des données manquantes, à symboliser par « ? ».

« t test » est le test paramétrique de comparaison de moyenne, qui suppose des distributions normales. « p(same mean) » est la probabilité que les deux moyennes soient identiques. « Sign test » et « Wilcoxon test » sont des tests non paramétriques de comparaison de médianes, qui ne supposent pas forcément que les distributions sont normales.

Dans tous les cas, « p(same median) » donne la probabilité que les médianes soient identiques, c'est à dire qu'il n'y ait pas de différence entre les populations étudiées (rappel : la médiane est la valeur pour laquelle la moitié des mesures sont plus grandes, et la moitié des mesures sont plus petites).

3.8 Normality (one sample) : test de normalité d'un échantillon

Pour une colonne sélectionnée, c'est un test de normalité de la distribution.



N est le nombre de cases mesurées, et pour 3 tests différents, p(normal) donne la probabilité que la distribution soit normale.

Vers le bas de la fenêtre, les quatre colonnes correspondent aux quatre quartiles, c'est à dire aux quatre parties de la distribution, normalement équiprobables:

- Upr limit est la valeur supérieure de l'intervalle
- Obs freq est le nombre d'individus observés dans cet intervalle (« fréquence absolue »)
- Exp freq est le nombre d'individus attendus dans cet intervalle.

3.9 Chi^2 : test du khi deux

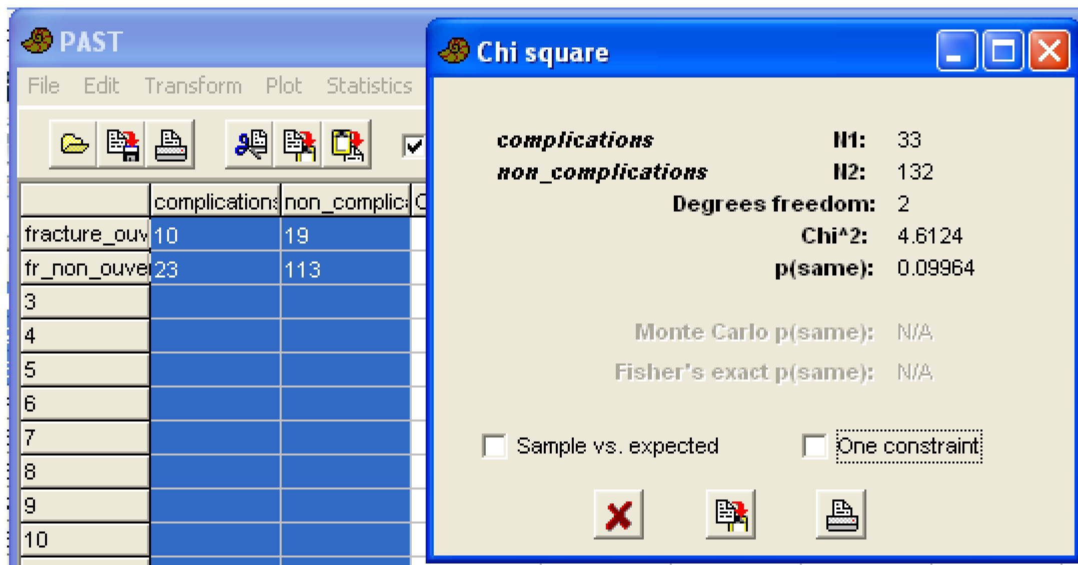
Ce test permet de comparer deux distributions (dans deux colonnes).

Pour chaque colonne, chaque case contient le nombre d'individus observés dans l'intervalle.

Ce test n'est valable que si les intervalles contiennent au moins cinq individus (sinon, il faut regrouper les intervalles).

p(same) indique la probabilité que les deux échantillons soient identiques (qu'il n'y ait pas de différences entre eux).

Pour le test du khi-deux, on suppose que chaque intervalle contient au moins 5 individus. Cette hypothèse n'est pas nécessaire pour les tests de Fisher (possible seulement si le nombre de lignes est 2) et Monte-Carlo.



Il faut cocher « Sample vs. expected » si la seconde colonne est une distribution théorique ; dans ce cas, on peut mettre des valeurs non entières dans cette seconde colonne.

La case « One constraint » doit être cochée s'il existe une contrainte supplémentaire, par exemple lorsqu'on met des pourcentages dans les cases, forcément le total de chaque colonne doit être égal à 100. Au contraire, lorsqu'on compare de nombres d'observations, il ne faut pas cocher cette case.

3.10 coefficient of variation : test d'égalité des coefficients de variation de deux échantillons

Le coefficient de variation vaut l'écart-type divisé par la moyenne de l'échantillon étudié. C'est un paramètre intéressant car il permet de comparer des mesures très hétérogènes, par exemple la taille d'une population d'éléphants et d'une population de souris. Il n'a aucune signification lorsqu'il peut exister des valeurs négatives et que la moyenne peut être nulle.

Cette option calcule les coefficients de variation de deux échantillons, et teste leur égalité.

3.11 Mann-Whitney : test U de comparaison de médianes, même si la distribution n'est pas normale

Il faut normalement que le nombre de mesures soit supérieur à 7, et que la forme des distributions soit semblable.



C'est un test « non paramétrique », testant si la médiane des deux distributions est semblable, utilisable même si les distributions ne sont pas normales (contrairement au test t, qui nécessite la normalité des distributions).

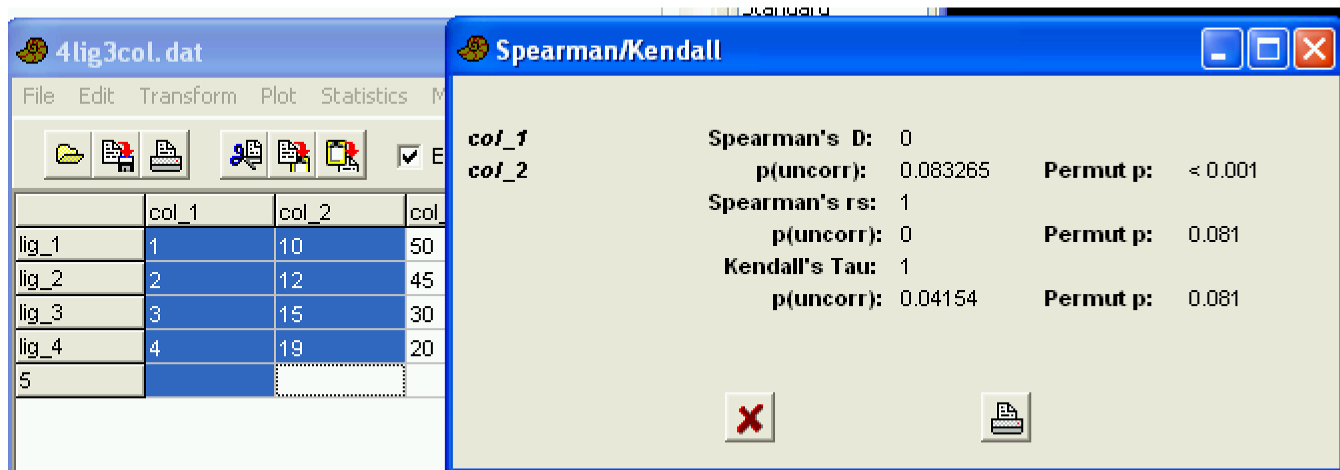
3.12 Kolmogorov Smirnov : teste si deux échantillons ont la même distribution

Après avoir sélectionné deux colonnes, ce test indique si les deux distributions sont différentes. C'est un test non paramétrique : il est utilisable avec des distributions de n'importe quel type (il n'y a pas besoin de supposer que les distributions sont normales).

Des valeurs manquantes sont possibles, en les indiquant par un point d'interrogation.

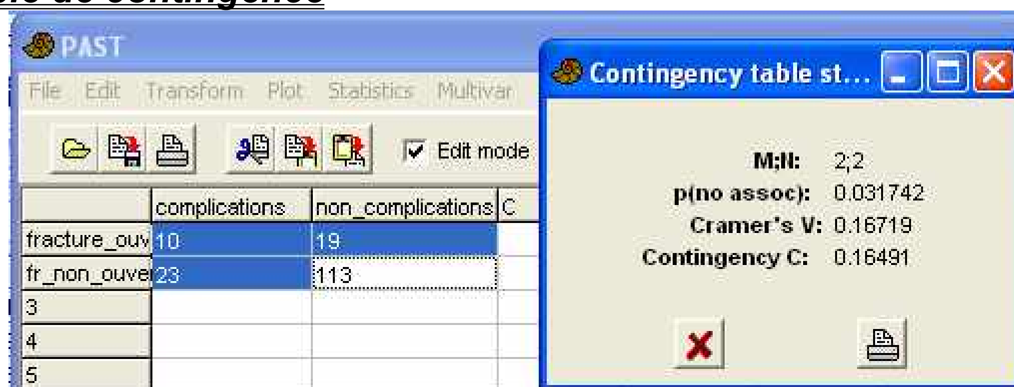
Si l'on veut simplement tester l'égalité des moyennes, il vaut mieux utiliser le test Mann-Whitney.

3.13 Spearman/Kendall : deux variables sont-elles corrélées ?



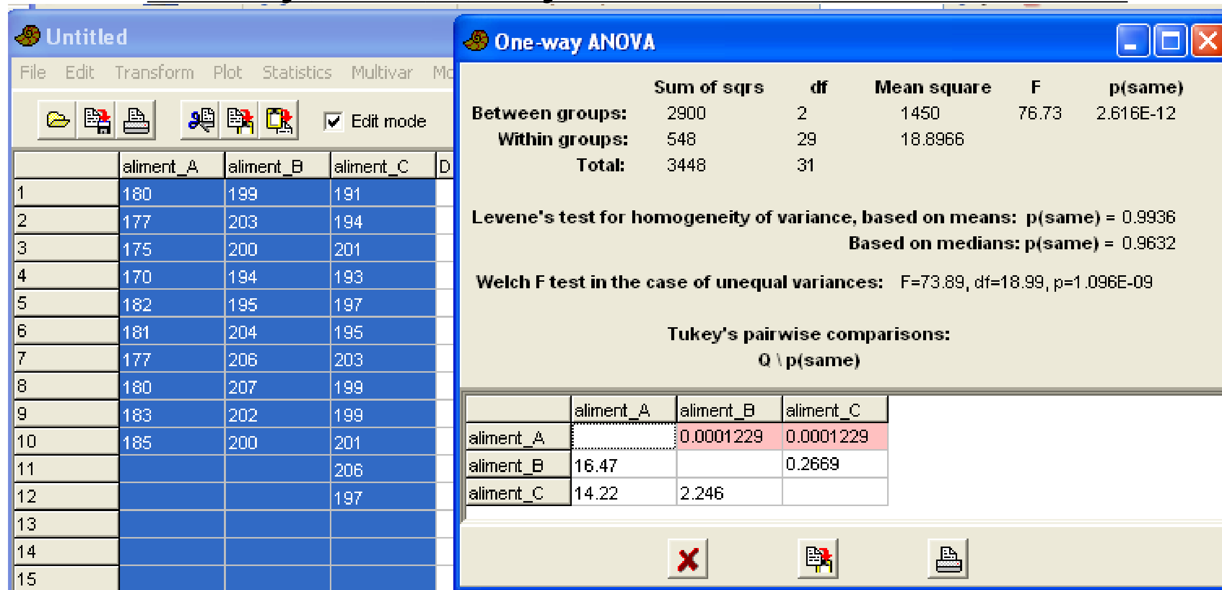
C'est aussi un groupe de tests non paramétriques, cherchant à savoir si deux variables sont corrélées. « p(uncorr) » indique la probabilité que les deux variables ne soient pas corrélées.

3.14 Contingency table : test d'indépendance de variables d'une table de contingence



Pour un couple de variables nominatives, on met dans les cases les nombres d'occurrences des diverses possibilités. Le test calcule la probabilité qu'il n'y ait pas de lien entre les variables, et l'indique par la valeur « p(no assoc) ».

3.15 One-way Anova = analyse de variance à un seul facteur



Fondamentalement, c'est une comparaison multiple de moyennes : les moyennes sont-elles différentes les unes des autres, ou bien les fluctuations entre les moyennes sont-elles simplement dues au hasard ? Il faut que les distributions des échantillons soient normales, que leurs effectifs et leurs variances soient semblables.

Les diverses colonnes correspondent aux échantillons étudiés (ici 3 : effet de 3 types d'aliments, A, B et C, sur le poids de poulets).

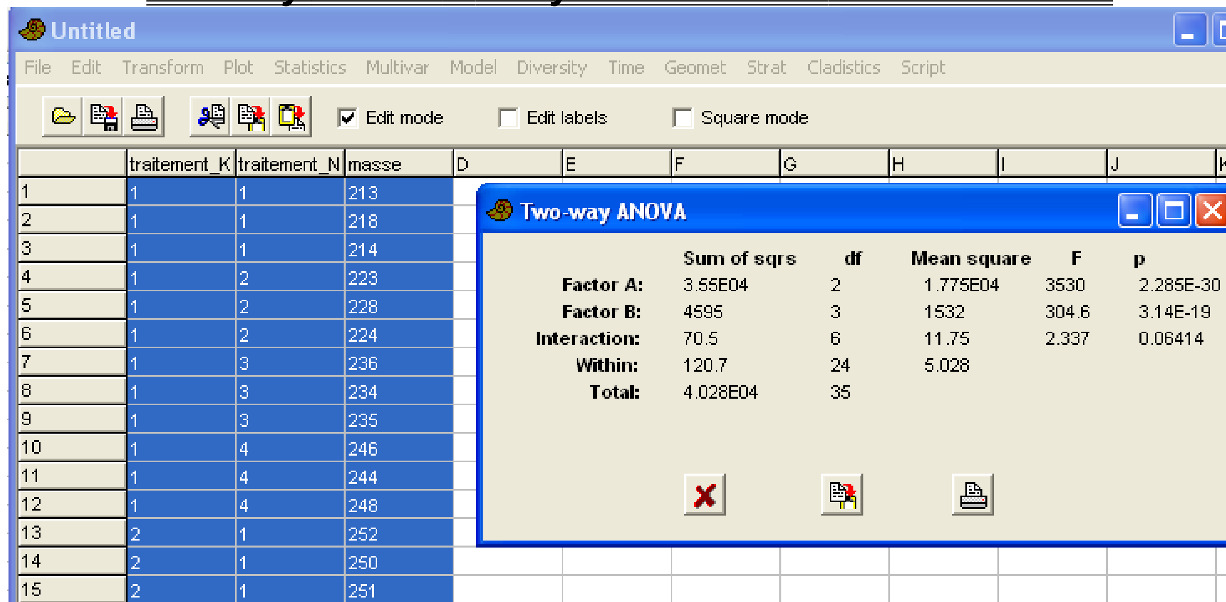
L'appel de cette fonction aboutit à une fenêtre :

- à gauche, la somme des carrés entre les groupes (intergroupe) et à l'intérieur des groupes (intragroupe), et le total (qui est la somme des deux précédentes).
- à droite, la probabilité de l'hypothèse nulle (les différences entre moyennes sont dues au hasard).
- Le test de Levene donne la probabilité que les variances soient homogènes, ce qui est nécessaire pour le test F précédent. Si les variances apparaissent différentes, il faut utiliser une autre méthode d'analyse de variance (Welch).

Une fois que l'analyse de variance a montré que les groupes étaient différents pour le caractère considéré, il reste à trouver pour quel(s) groupe(s) les différences sont importantes.

Dans le triangle supérieur sont indiquées les probabilités que les échantillons aient une moyenne égale : A est franchement différent de B et C (car la probabilité d'égalité est très faible), alors que B et C ne sont pas très différents (probabilité d'égalité 0,2669).

3.16 Two-way ANOVA : analyse de variance à deux facteurs



De même que pour l'analyse de variance à un facteur, il faut que les distributions soient normales,et que les variances et effectifs soient semblables.

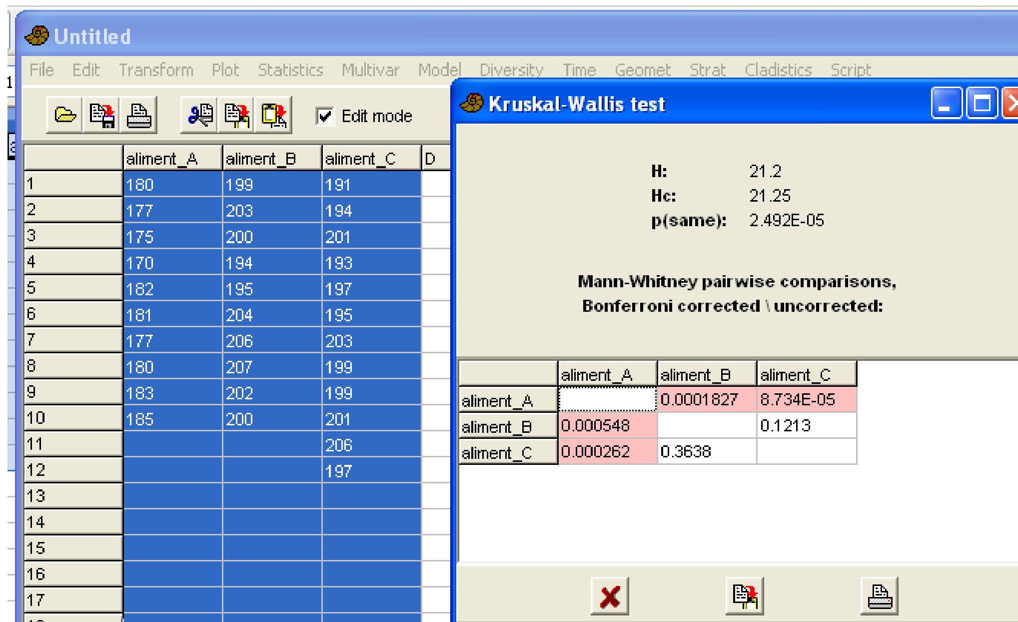
Une première colonne contient les modalités du premier facteur (noté facteur A), codées de 1 à n. La deuxième colonne contient les modalités du deuxième facteur (noté facteur B), codées de 1 à m.

Cette option donne l'effet du facteur A, du facteur B et de l'interaction entre les deux facteurs.

La dernière colonne donne la probabilité que cet effet soit simplement au hasard. Ici, la probabilité que les différences soient dues au hasard est très faible : il est quasiment certain que le facteur A a un effet, très fortement probable que le facteur B a un effet, et assez probable que l'interaction entre les deux existe (puisque la probabilité que l'interaction soit due au hasard est de 0,06414.

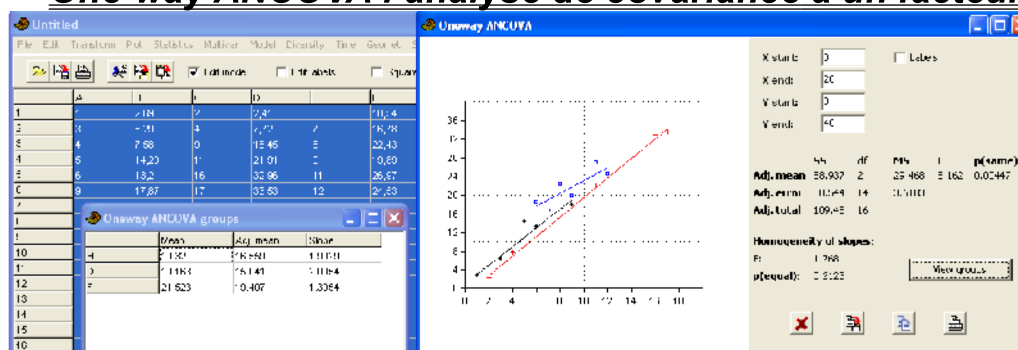
3.17 Kruskal-Wallis : comparaison multiple de médianes, par une sorte d'analyse de variance où les distributions ne sont pas forcément normales

C'est une alternative non paramétrique à l'analyse de variance, où on compare les médianes de quelques groupes univariés (dans les colonnes). On ne suppose pas que les distributions sont normales, mais il faut que les distributions aient la même forme entre les divers groupes.



Là aussi, « p(same) » indique la probabilité que les distributions soient les mêmes, et la matrice contient les probabilités que les couples de distributions soient identiques.

3.18 One-way ANCOVA : analyse de covariance à un facteur



Cette méthode étudie l'effet d'une variable qualitative (facteur) et d'une variable continue (covariable) sur une variable réponse.

On doit avoir quelques paires de colonnes correspondant aux différentes modalités de la variable qualitative. Chaque paire de colonne correspond à un ensemble de données plus ou moins corrélées (ici 3 paires).

L'analyse de covariance indique si les pentes peuvent être considérées comme différentes ou non, en plus de la comparaison multiple de moyennes.

Ici, la probabilité que les moyennes soient identiques pour les trois groupes est faible (0,00447), on peut donc considérer que les moyennes sont différentes.

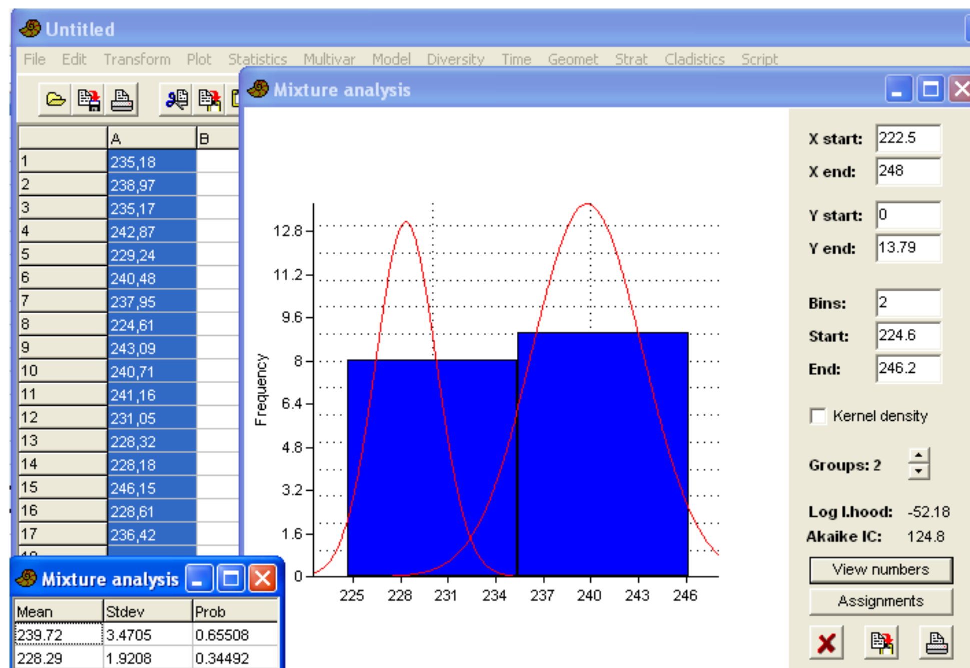
Par contre, la probabilité que les pentes soient homogène est assez forte (0,2123) : on ne peut donc pas considérer que les pentes sont différentes.

Le bouton « View groups » fait apparaître une fenêtre supplémentaire, où pour chaque variable dépendante (la 2e colonne de chaque paire), la moyenne calculée, la moyenne calculée par ajustement pour l'ensemble des valeurs, et la pente de la droite de régression sont affichées.

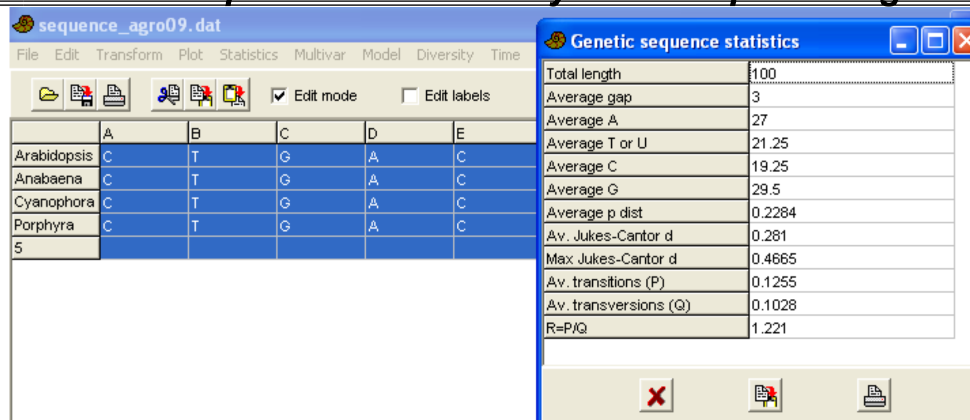
3.19 Mixture analysis = analyse des mélanges, pour une population hétérogène

C'est une méthode utilisée lorsqu'on a en mélange deux ou quelques populations, par exemple

lorsqu'on a un mélange de mâles et de femelles. Chaque population est à peu près normale, mais le mélange des diverses populations aboutit à une distribution non-normale, par exemple avec deux ou plusieurs modes.



3.20 Genetic sequence stats : analyse de séquences génétiques

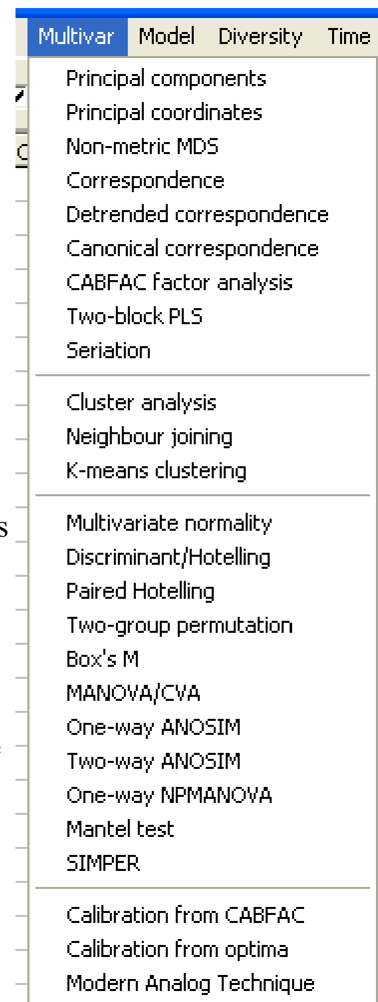


Cette analyse statistique n'est possible que pour des séquences de nucléotides codés ATCG, ou 1234.

Elle donne :

- la longueur totale de la séquence analysée,
- le nombre moyen de données manquantes pour chaque ligne,
- la quantité moyenne de chaque nucléotide,
- les distances moyennes entre chaînes selon deux méthodes,
- le nombre moyen de transitions (remplacement d'une base purique par une autre, ou d'une base pyrimidique par une autre) et de transversions (remplacement d'une base purique par une base pyrimidique, ou inversement). Normalement, les transitions sont plus fréquentes que les transversions.
- le rapport transition/transversion.

4 Multivar : statistiques multivariables



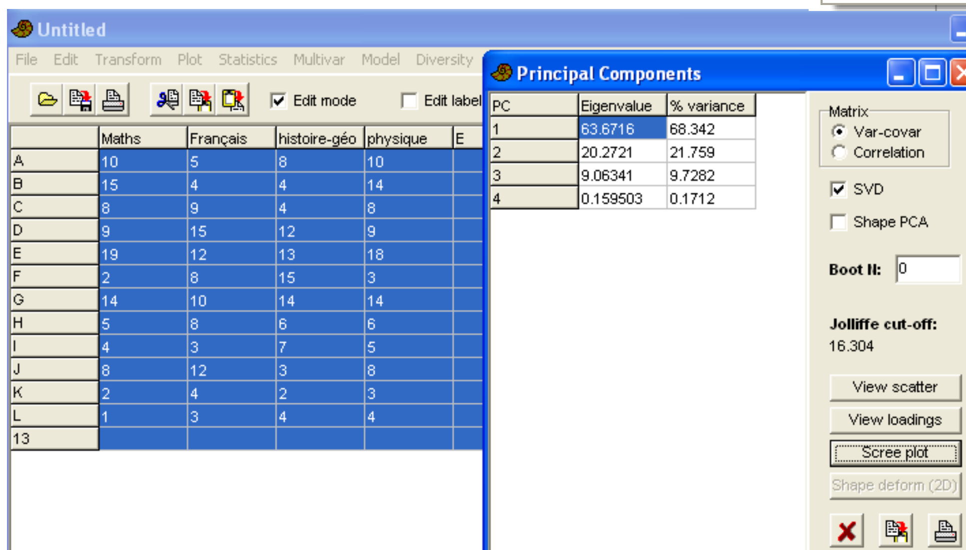
Ce menu regroupe diverses fonctions permettant de manipuler des tableaux avec de nombreuses variables, et d'en extraire l'information sous une forme plus simple. D'une façon générale, les colonnes correspondent aux diverses variables, et les lignes aux différents individus (lieux, points de mesure, etc.) où ont été mesurées ces variables.

Les deux méthodes les plus importantes à connaître sont l'analyse en composantes principales, pour les données de mesures, et l'analyse des correspondances, pour les données de comptages.

4.1 Principal components analysis : Analyse en composantes principales (ACP = PCA)

C'est une technique qui permet de récapituler en deux ou quelques variables synthétiques, les « composantes principales », la majorité de l'information contenue dans les variables initiales

L'exemple ci-dessus correspond aux notes de 12 élèves dans quatre matières : Math, Français, histoire-géo, physique. Ces quatre matières sont résumées par quatre composantes, mais ce sont les deux ou trois premières qui sont les plus intéressantes, car elles possèdent la plus grande partie de l'information initiale.



	Maths	Français	histoire-géo	physique
A	10	5	8	10
B	15	4	4	14
C	8	9	4	8
D	9	15	12	9
E	19	12	13	18
F	2	8	15	3
G	14	10	14	14
H	5	8	6	6
I	4	3	7	5
J	8	12	3	8
K	2	4	2	3
L	1	3	4	4
13				

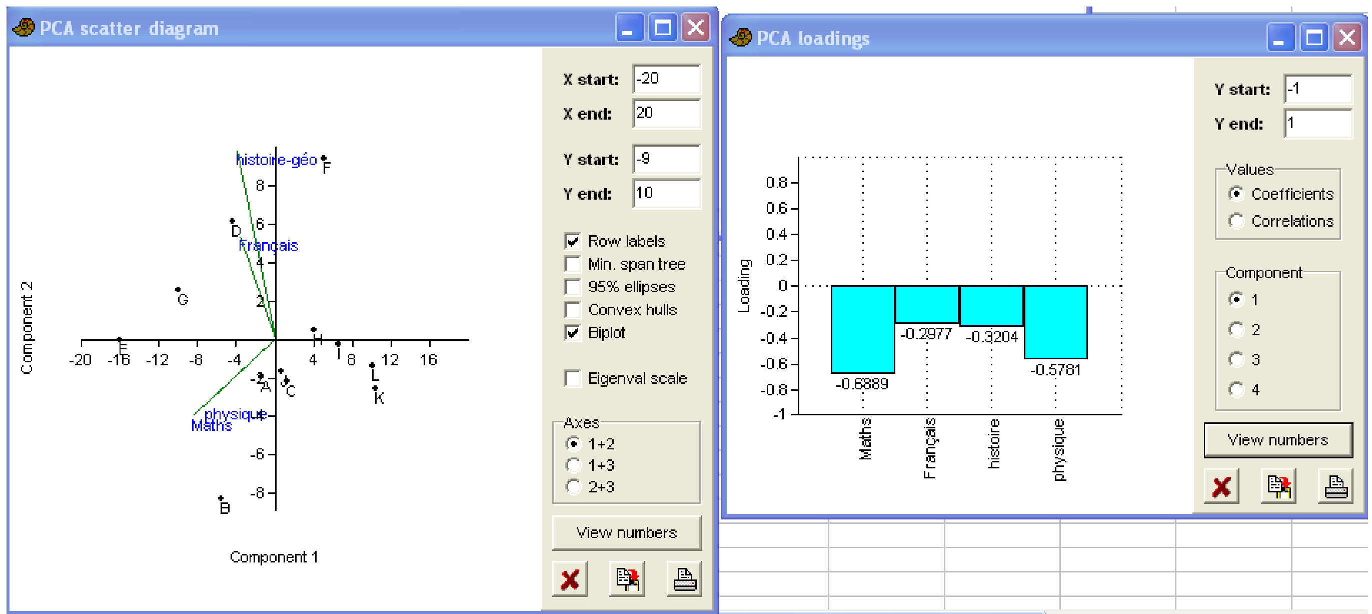
PC	Eigenvalue	% variance
1	63.6716	68.342
2	20.2721	21.759
3	9.06341	9.7282
4	0.159503	0.1712

Le déclenchement de l'option ouvre une fenêtre donnant en particulier le pourcentage de variance initiale expliqué par chaque composante. La première donne la majorité de l'information (68,342%), la seconde donne aussi un peu d'information (20,272%), mais les suivantes beaucoup moins. « Eigenvalue » est la valeur propre de la composante en question.

Techniquement, l'analyse en composantes principales peut se faire soit à partir de la matrice des variances-covariances (option par défaut), soit à partir de la matrice des corrélations entre variables (option à cocher, à choisir si les variables sont dans des unités très différentes).

« Jolliffe cut-off » donne la valeur-seuil de la valeur propre à partir de laquelle on peut considérer que la composante a de l'importance. Ici, le seuil est 16,304, ce qui signifie que les composantes 3 et 4 peuvent être négligées, et que l'on ne gardera que les composantes 1 et 2.

En mettant un nombre supérieur à 1 dans la ligne « Boot N », on obtient deux colonnes supplémentaires, indiquant un intervalle de confiance pour les valeurs propres. Les cases « SVD » et « Shape PCA » ne sont actives que dans ce cas.



Le bouton « View scatter » provoque l'apparition de graphiques montrant les individus (et éventuellement les variables, en cochant « Biplot »), dans un couple de composantes principales (axes 1-2, ou 1-3, ou 2-3, mais dans la pratique, c'est le couple 1-2 qui contient le maximum d'informations). Le bouton « View numbers » provoque l'apparition des coordonnées des individus dans les quatre composantes principales.

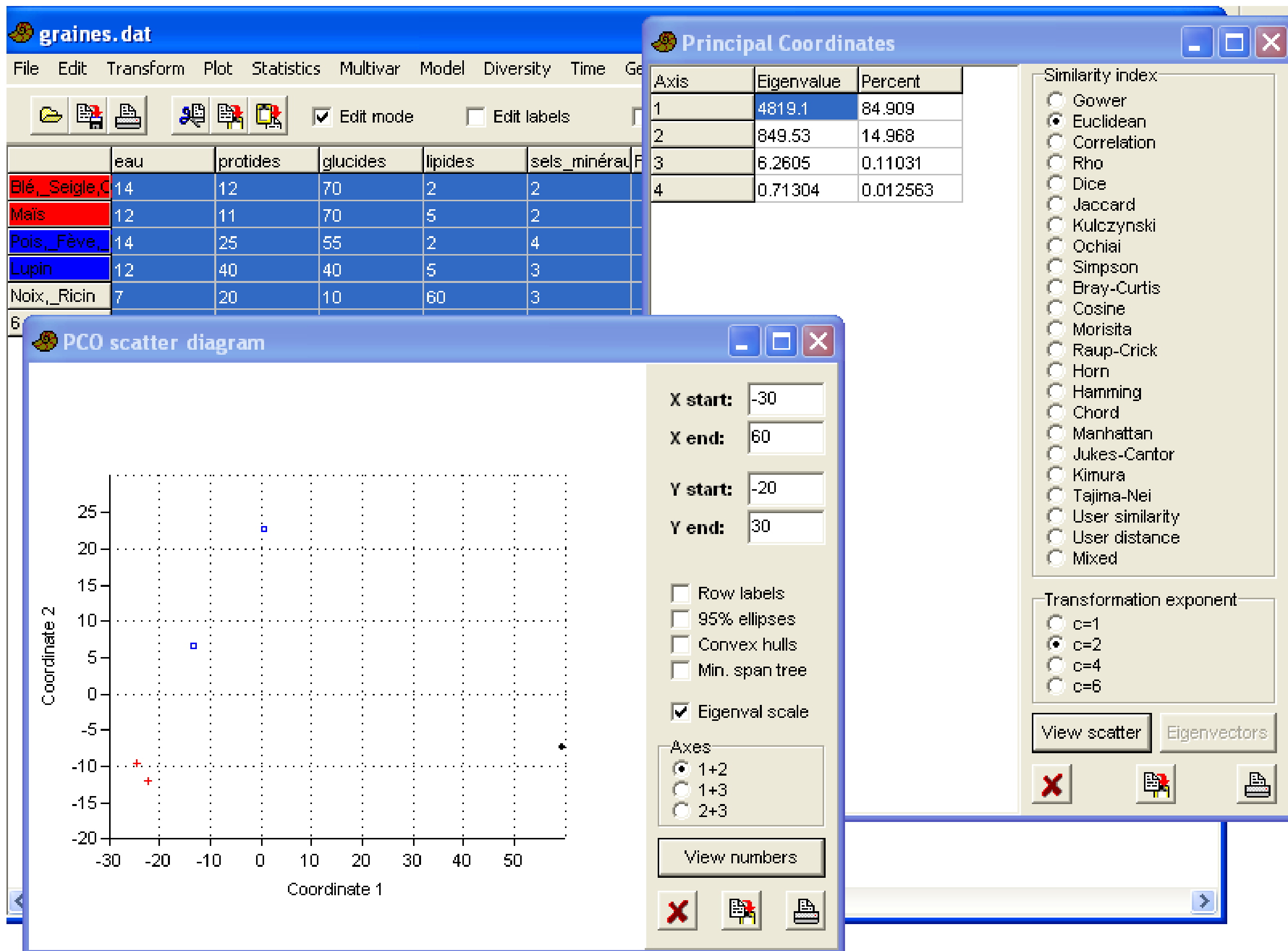
Le bouton « View loadings » provoque l'apparition d'une fenêtre montrant le poids des diverses variables initiales pour la détermination d'une composante principale.

« Scree plot » trace un graphique des valeurs propres pour toutes les composantes principales.

Si des groupes de coloration ont été définis précédemment (« Edit | Row color/symbol »), ces couleurs sont visibles sur le graphique obtenu par « View scatter ». Ceci peut permettre de visualiser divers groupes dans diverses régions de ce graphe.

4.2 Principal coordinates : analyse en coordonnées principales, proche de l'ACP

C'est une extension de l'analyse en composantes principales, qui permet d'utiliser des distances variées, autres que la simple distance euclidienne.



A partir des mesures, le logiciel trouve les valeurs propres de la matrice de distance entre les points. Si les lignes ont été colorées (menu « Edit »), les différents groupes de couleurs sont visibles sur le graphique

4.3 Non-metric MDS : positionnement multidimensionnel non métrique

Cette technique part aussi d'une matrice de similarités entre les individus. Après avoir choisi cette option, il faut choisir une méthode de calcul des similarités. Après un petit moment d'attente, le graphe est tracé, où les divers individus sont positionnés selon deux axes..

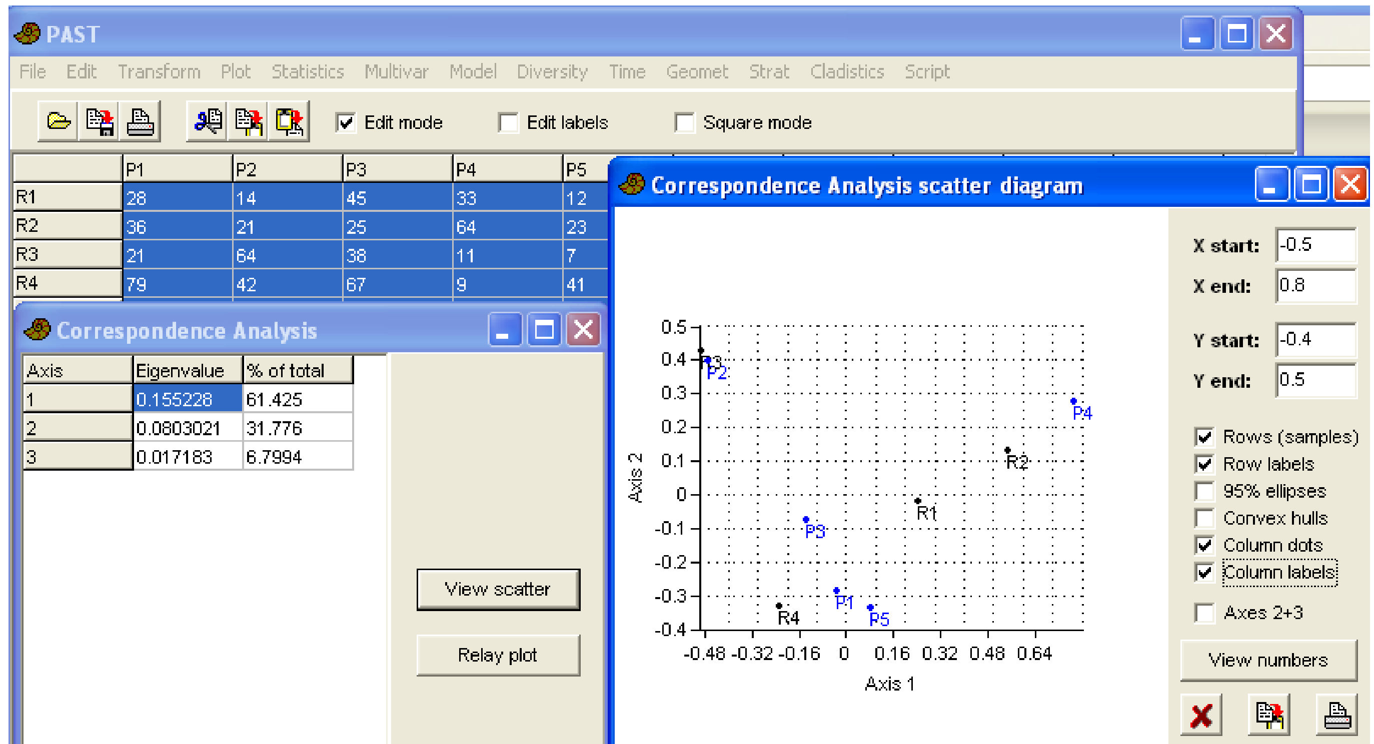
Le bouton « Shepard plot » ouvre une nouvelle fenêtre :plus les points sont alignés, meilleure est l'analyse.

4.4 Correspondence = analyse des correspondances = analyse factorielle des correspondances (CA = AFC)

C'est une méthode assez semblable à l'analyse en composantes principales, mais pour des valeurs de comptages.

On part d'un tableau de contingence, c'est à dire d'un tableau de nombres (entiers, en principe), où les lignes représentent les modalités d'un facteur, et les colonnes les modalités d'un autre facteur. Le contenu des cases correspond au nombres d'occurrences de la combinaison des deux facteurs.

Un exemple classique est en écologie végétale : n espèces de plantes ont été observées dans m lieux. On obtient ainsi une matrice des n x m, et l'analyse des correspondances permet d'associer certaines espèces à certains lieux.



4.5 Detrended correspondance analysis = analyse des correspondances redressée

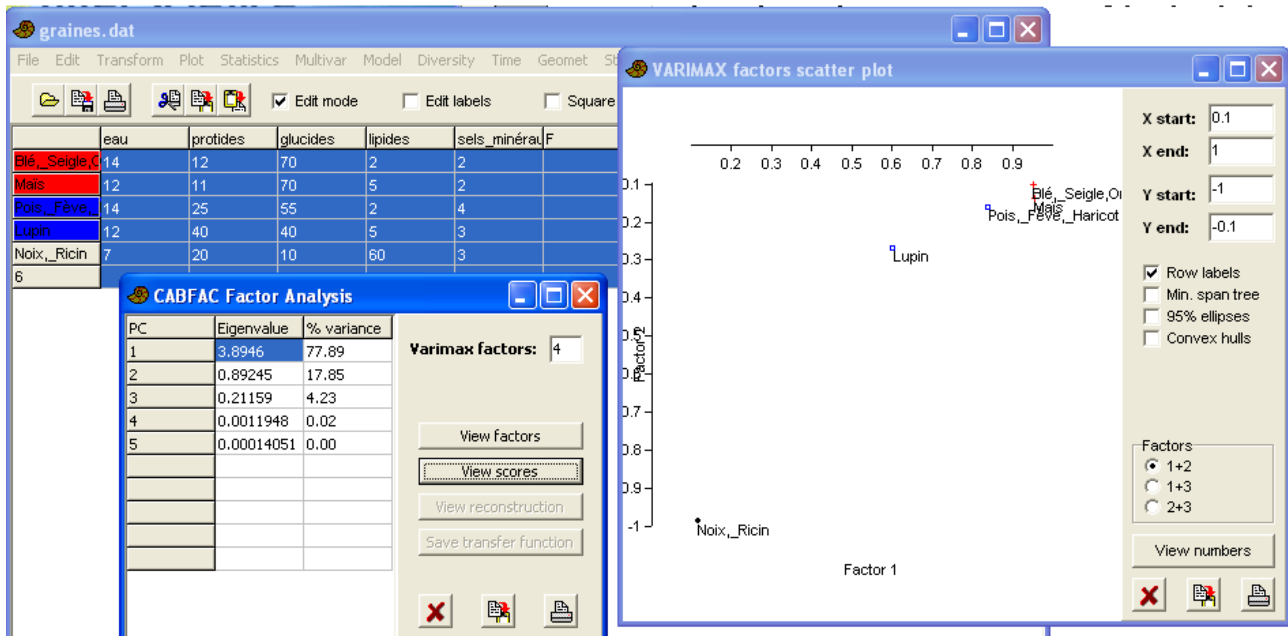
C'est une variante de l'analyse des correspondances.

4.6 Canonical correspondance analysis = analyse canonique des correspondances

Elle permet par exemple d'étudier les facteurs écologiques qui expliquent les répartitions des espèces vivantes.

4.7 CABFAC factor analysis

C'est une méthode factorielle d'analyse des données de comptage, éventuellement en association avec des données environnementales.



La première colonne peut contenir des données environnementales numériques, par exemple la température. Dans ce cas, on peut sauver dans un fichier une « fonction de transfert » indiquant en particulier les coefficients de régression pour la première colonne.

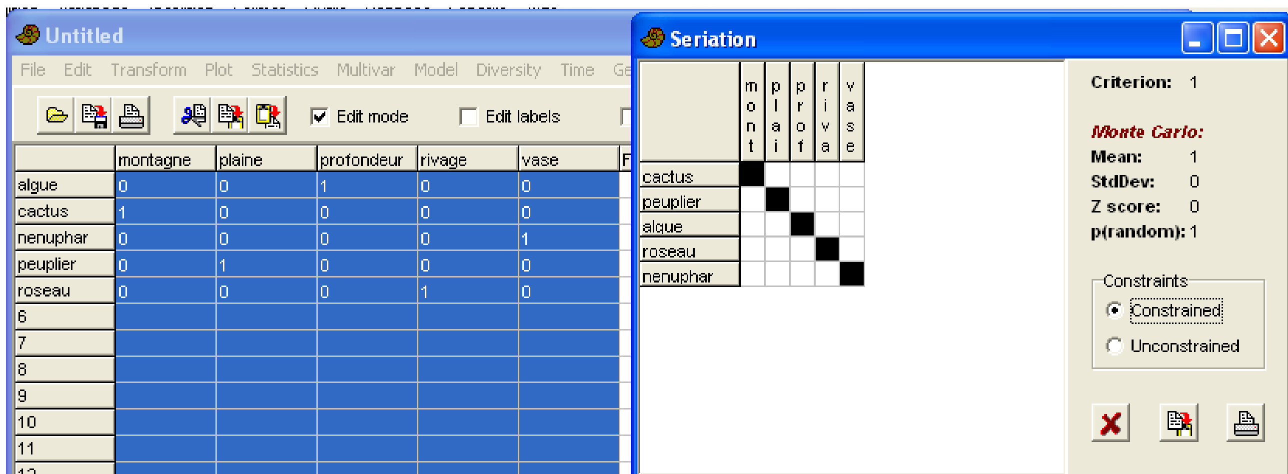
4.8 *two blocks PLS : Moindre carrés partiels sur deux blocs*

Les colonnes doivent être constituées de données continues (mesures), d'abord le premier bloc, puis le deuxième bloc, par exemple d'abord les données morphométriques, puis ensuite les données environnementales.

C'est une méthode comparable à l'analyse en composantes principales, mais qui étudie la covariation entre les deux blocs de variables.

4.9 *Seriation*

C'est une méthode qui utilise une matrice de présence/absence, avec par exemple les espèces en lignes et les milieux en colonnes. Elle essaie de réorganiser les lignes et les colonnes de façon à concentrer les présences autour de la diagonale.

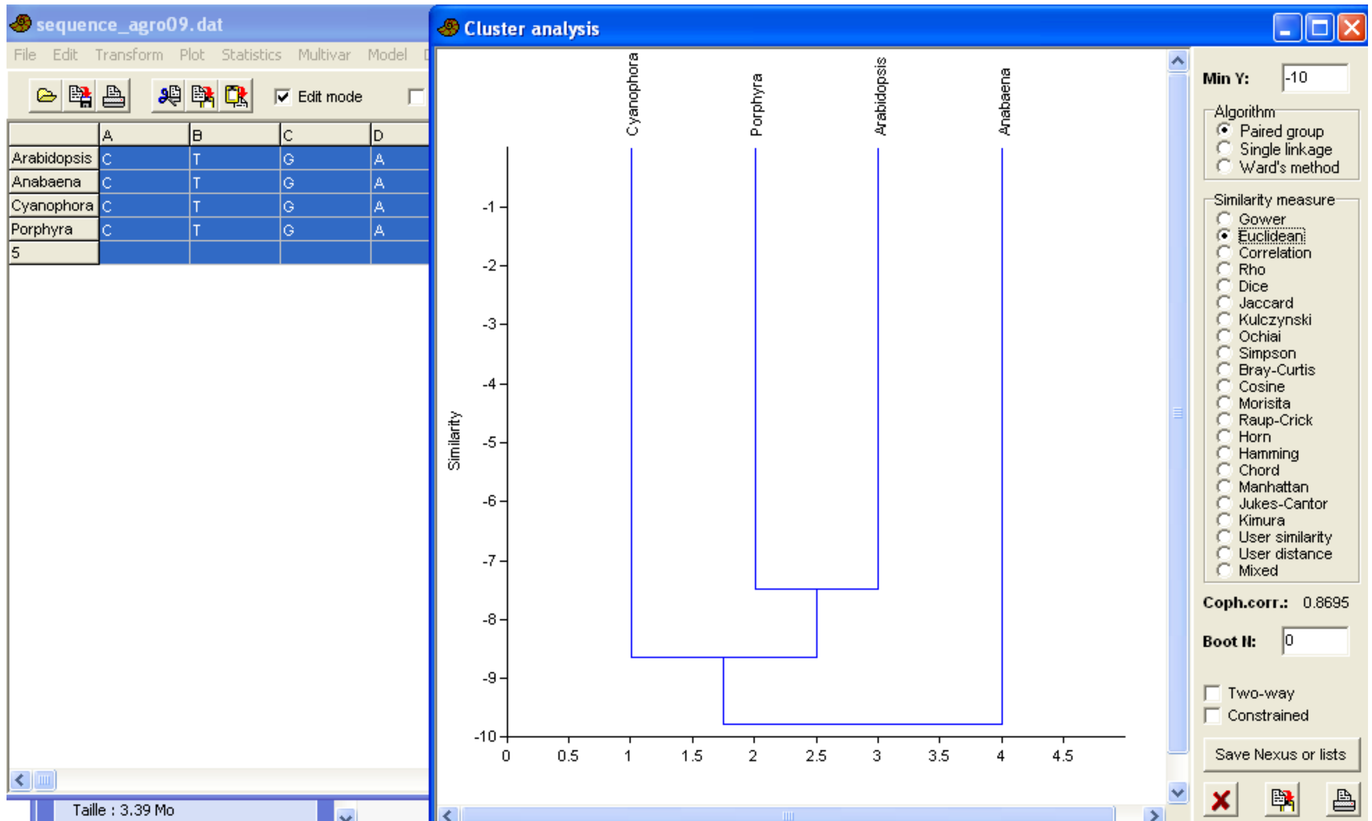


Il y a deux méthodes d'optimisation :

- « constrained » = optimisation contrainte : seules les lignes sont mobiles, et les colonnes doivent rester fixes
- « unconstrained » = optimisation non contrainte : lignes et colonnes sont mobiles, de façon à concentrer les présences encore plus autour de la diagonale.

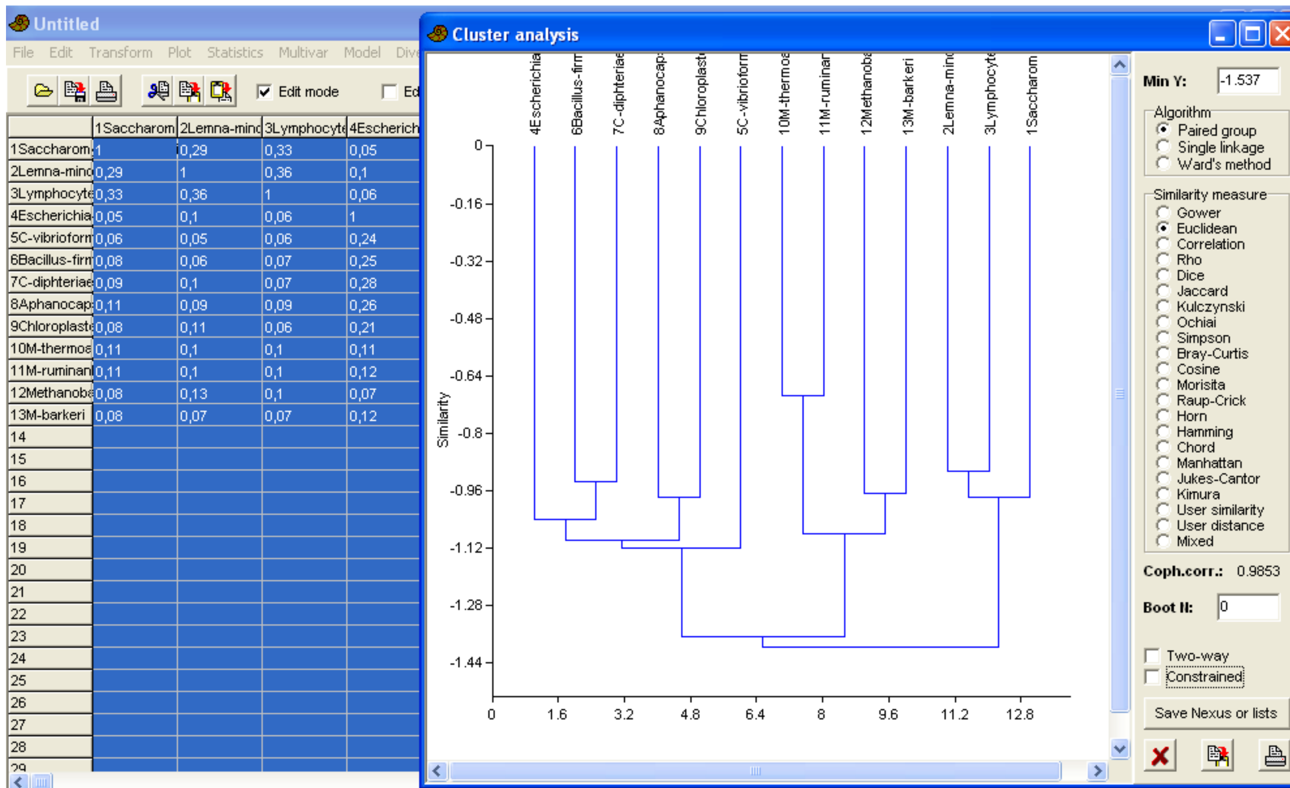
4.10 Cluster analysis = Regroupements en arbres

Cette analyse réalise des dendrogrammes, ou graphiques en arbres : les observations sont reliées entre elles par des branches, d'autant plus courtes que les observations sont semblables. C'est en particulier très utilisé en biologie de l'évolution, pour reconstituer des arbres phylogénétiques.



Chaque ligne correspond à un individu (ou une espèce, ou un lieu de mesure...), et chaque colonne correspond à un caractère, pouvant exister sous plusieurs modalités ou plusieurs valeurs.

On peut aussi partir d'une matrice carrée de différences ou de ressemblances :



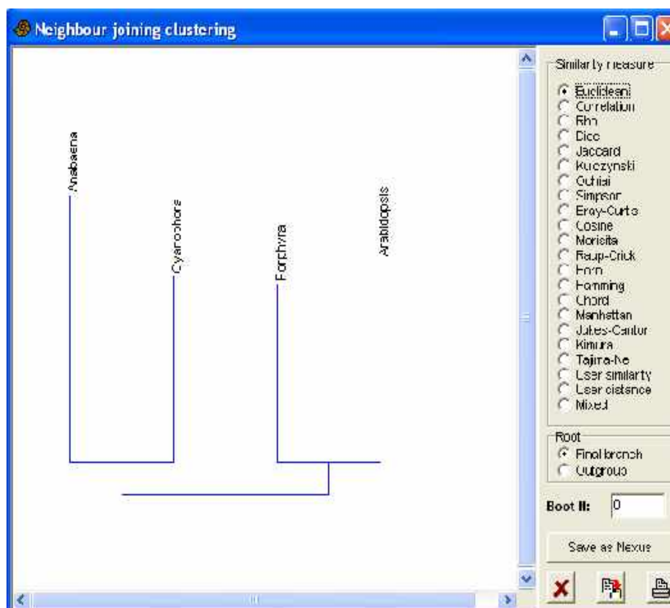
Cette option regroupe les individus par leur ressemblance (selon diverses méthodes possibles).

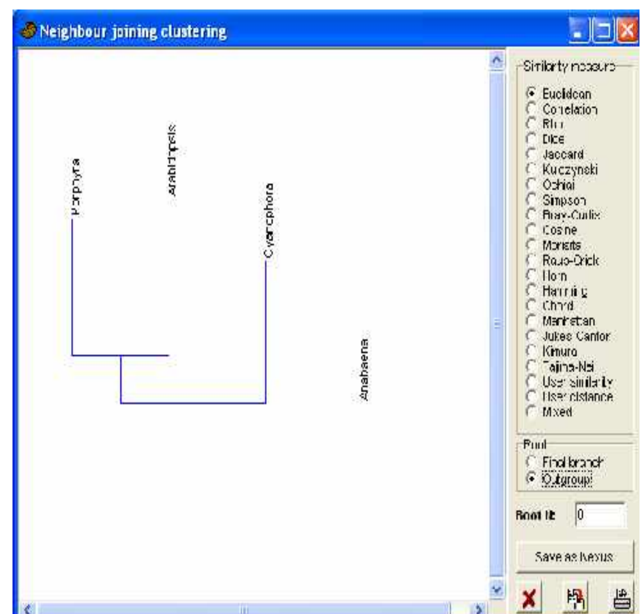
Cocher « Two-way » fait apparaître deux arbres : d'une part l'arbre des colonnes tracé grâce aux lignes, d'autre part l'arbre des lignes tracé grâce aux colonnes.

4.11 Neighbour joining

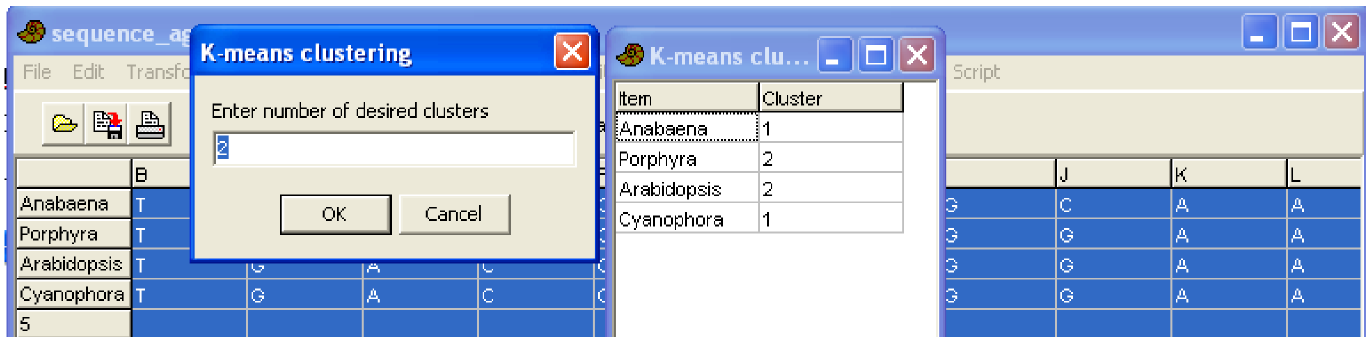
C'est aussi une méthode de regroupement en arbres.

Si l'on peut définir un extragroupe, il faut le mettre en première ligne pour qu'il soit considéré comme tel lorsqu'on coche « outgroup » au lieu de « final branch ».





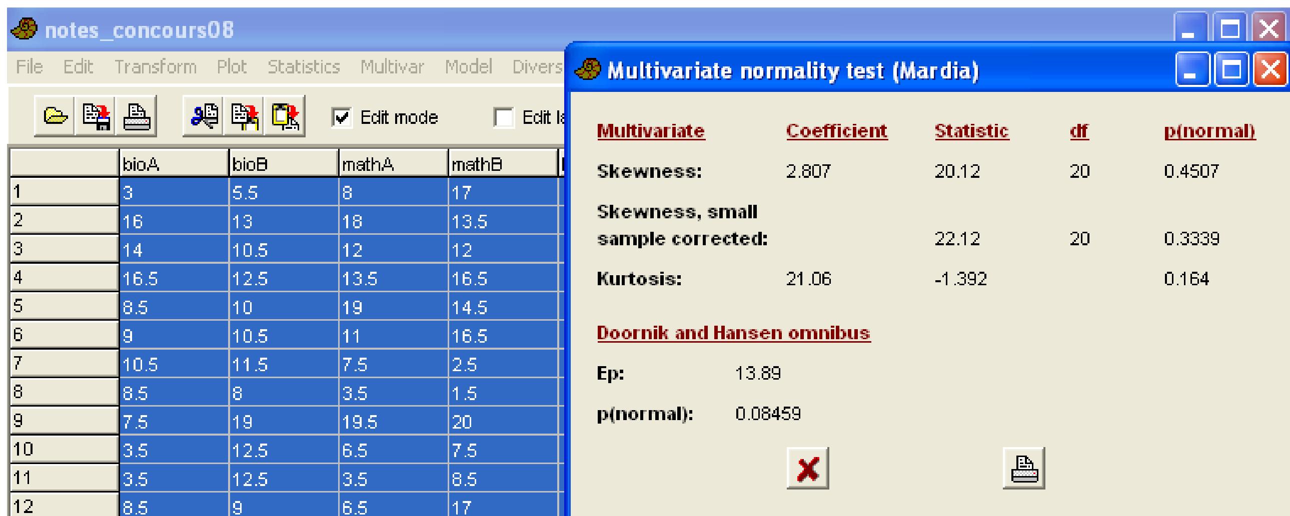
4.12 K-means clustering



Par des méthodes voisines des précédentes, les individus sont regroupés en paquets (« clusters »). On indique le nombre de paquets souhaités, et le logiciel range tous les individus semblables dans un paquet, selon leur proximité d'après les variables étudiées.

4.13 Multivariate normality = test de normalité sur plusieurs variables

Les données sont un ensemble de colonnes contenant des valeurs numériques (continues, comme pour une analyse en composantes principales).



Cette option teste la normalité de l'ensemble des variables. La partie supérieure teste l'asymétrie (Skewness) et l'aplatissement (Kurtosis) des distributions ; p(normal) indique la probabilité de la normalité (asymétrie et aplatissement nuls).

La partie inférieure est un test de normalité.

4.14 Discriminant / Hotelling = Analyse discriminante , et test T² de Hotelling

On suppose des variables numériques continues, pour des individus (points de mesure) marqués de différentes couleurs (menu « Edit »), de distribution normale, et de covariances égales.

Ce test indique si les deux groupes doivent être considérés comme différents ou non.

4.15 Paired Hotelling = test T² de Hotelling pour données appariées

C'est l'équivalent du précédent, mais où les individus des deux groupes sont appariés. Il faut que les individus de chaque groupe soient consécutifs. Le premier individu du premier groupe est considéré comme apparié au premier du deuxième groupe, le deuxième apparié au deuxième, etc.

4.16 Two-group permutation : test de permutation pour deux groupes multivariés

Là encore, il faut deux groupes de données multivariées (valeurs numériques continues, comme pour une ACP), marqués par deux couleurs différentes. Les lignes de chaque groupe doivent être consécutives.

Ce test est comparable au test de Hotelling : il indique si les deux groupes peuvent être considérés comme équivalents, ayant des moyennes égales.

4.17 Box's M = test d'égalité des matrices de covariances pour deux groupes de données

Comme pour les tests précédents, on utilise des données réparties en deux groupes, de couleurs différentes (menu « Edit »). Ce test indique si les matrices de variances-covariances des variables peuvent être considérées comme égales pour les deux groupes.

4.18 MANOVA/CVA = analyse de variance multiple / analyse canonique des variables

Là encore, il faut deux ou plusieurs groupes d'individus (points de mesure) marqués de couleurs différentes ; il faut que le nombre de ces points de mesures soit supérieur au nombre des variables.

4.19 One-way ANOSIM = analyse de similarités

Il faut encore deux ou plusieurs groupes de données multivariées, où chaque groupe est caractérisé par une couleur différente. Cette option teste les différences entre les groupes, qui sont supposés avoir même médiane et étendue (c'est un test non paramétrique : on ne suppose pas que les distributions sont normales). Il est fondé sur l'analyse des distances entre les groupes par rapport aux distances à l'intérieur des groupes.

4.20 Two-way ANOSIM = analyse de similarités à deux facteurs

C'est le même principe que l'analyse des similarités à un facteur (option précédente), mais pour deux facteurs.

Les données commencent par deux colonnes, correspondant aux deux facteurs codés par des entiers. Les colonnes suivantes sont les données multivariées.

4.21 One-way NPMANOVA = analyse de variance multiple non paramétrique à un facteur

Il faut des données groupées en deux ou quelques groupes, indiqués par des couleurs différentes (menu « Edit | Row color/symbol »), où les groupes ont des distributions semblables.

4.22 Mantel test : test de corrélation entre matrices de distance

Il faut deux groupes de données multivariées, avec des couleurs différentes, ou bien deux matrices symétriques de distances ou de similarités.

Il teste les corrélations entre les matrices de distances

4.23 SIMPER = pourcentage de similarité

Il faut un tableau de deux ou plusieurs groupes d'individus, marqués par des couleurs, et des variables de comptage. On suppose que les groupes sont indépendants.

4.24 Calibration from CABFAC

Normalement, si on a sauvé dans un fichier une fonction de transfert lors de l'opération « CABFAC factor analysis », on doit pouvoir l'utiliser pour prédire la variable d'environnement à partir de données d'observation.

4.25 Calibration from optima

On part d'une matrice d'abondance, avec les lignes correspondant aux échantillons, et les colonnes correspondant aux taxons, mais avec les trois premières lignes particulières :

- première ligne : l'optimum du taxon
- deuxième ligne : les tolérances du taxon
- troisième ligne : abondance maximale du taxon
- (et lignes suivantes : comptages d'abondance du taxon dans les divers lieux)

Les trois premières lignes peuvent être obtenues par le menu « Model », où il existe une option « Abundance » et une « Species packing ».

4.26 Modern Analogue Technique

Là aussi, on part d'une matrice d'abondance, avec les taxons en colonnes et les observations en lignes. Les premières lignes correspondent aux taxons actuels, dont on peut connaître les facteurs écologiques, et les dernières lignes correspondent aux taxons fossiles.

La première colonne contient des données environnementales pour les taxons actuels, et un point d'interrogation pour les taxons fossiles.

5 Model : modélisation

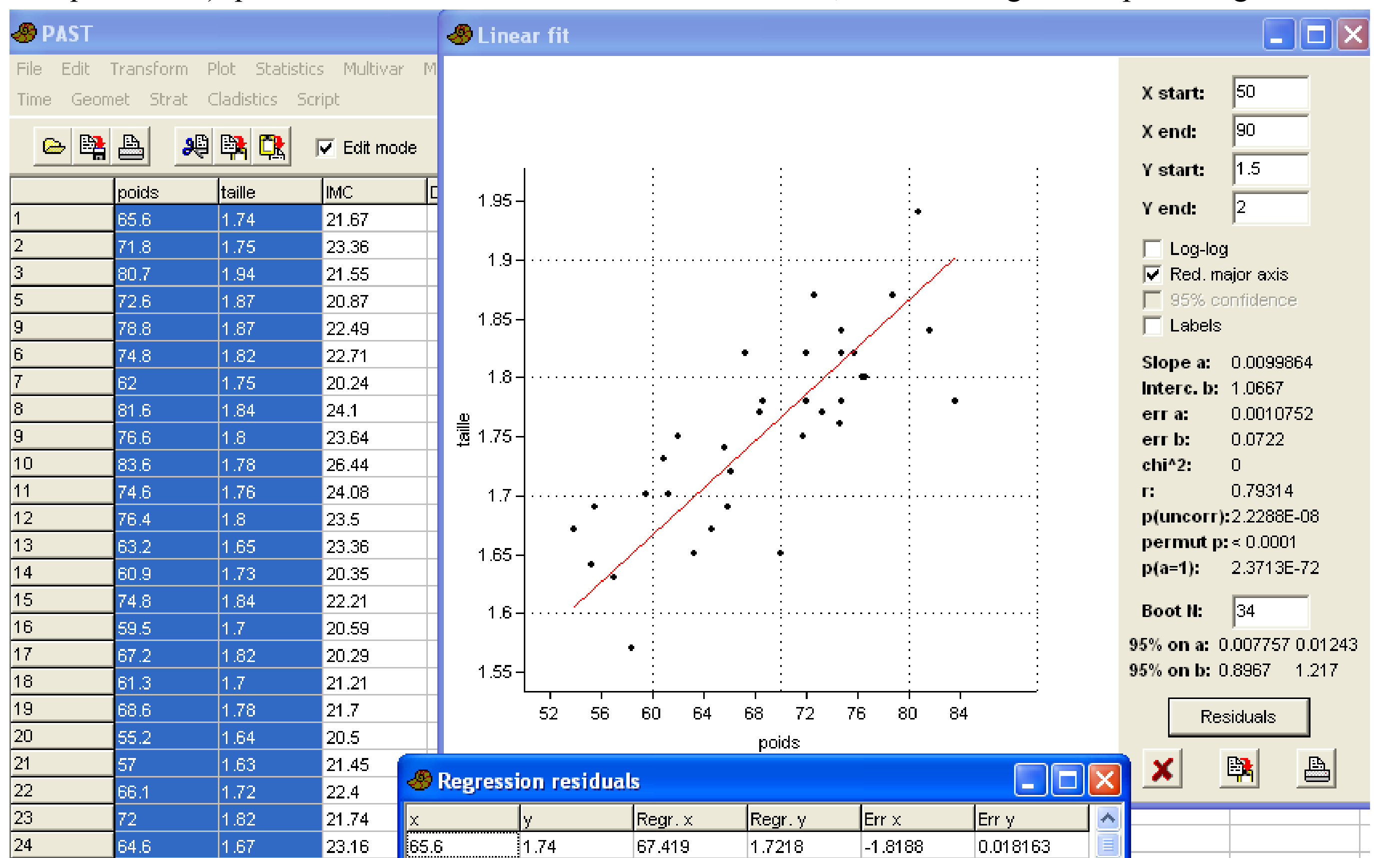
Model Diversity Time Ge
Linear
Linear 1 indep, n dep
Linear n indep, 1 dep
Sinusoidal
Polynomial
Logistic
Smoothing spline
Abundance
Species packing

A partir de valeurs numériques d'observation, on essaie de calculer une formule de prédiction.

5.1 Linear : régression linéaire

La régression linéaire est le modèle le plus simple.

Attention ! Ici, par défaut, ce n'est pas la régression standard (où la régression de y en x est d'autant plus différente de la régression de x en y que le nuage est étalé), mais la « Reduced Major Axis » (où la régression de x en y et la régression de y en x sont équivalentes) qui sert de base aux calculs. En d'autres termes, la droite rouge correspond au grand



axe de symétrie du nuage de points.

Pour obtenir la régression standard, qui est la plus utilisée en statistiques, décocher la case Red major axis. Dans le cas de cette régression standard, on peut tracer le domaine de confiance à 95% de la droite de régression, alors que ce n'est pas possible avec la régression RMA.

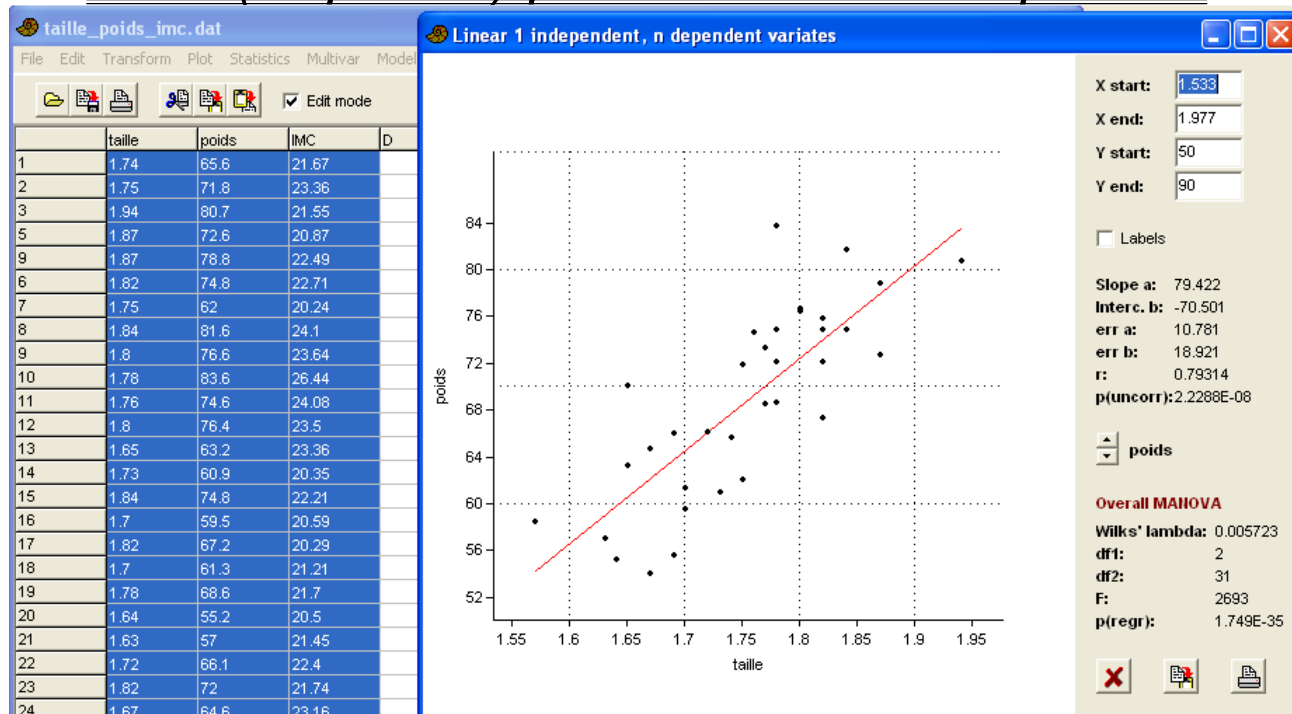
Slope a désigne la pente de la droite, Interc b désigne l'ordonnée à l'origine (différents selon le mode de régression).

r est le coefficient de corrélation linéaire, et p(uncorr) est la probabilité qu'il n'y ait pas de corrélation entre les données (ces deux valeurs sont identiques pour la régression standard et la régression RMA).

Cocher la case « Log-log » provoque le traçage du graphe ln(y) = f(ln(x))

Ne pas oublier qu'une exponentielle peut être modélisée par régression linéaire après transformation des valeurs Y en logarithme.

5.2 Linear 1 indep, n dep = régression linéaire multiple pour une variable (indépendante) qui détermine n variables dépendantes

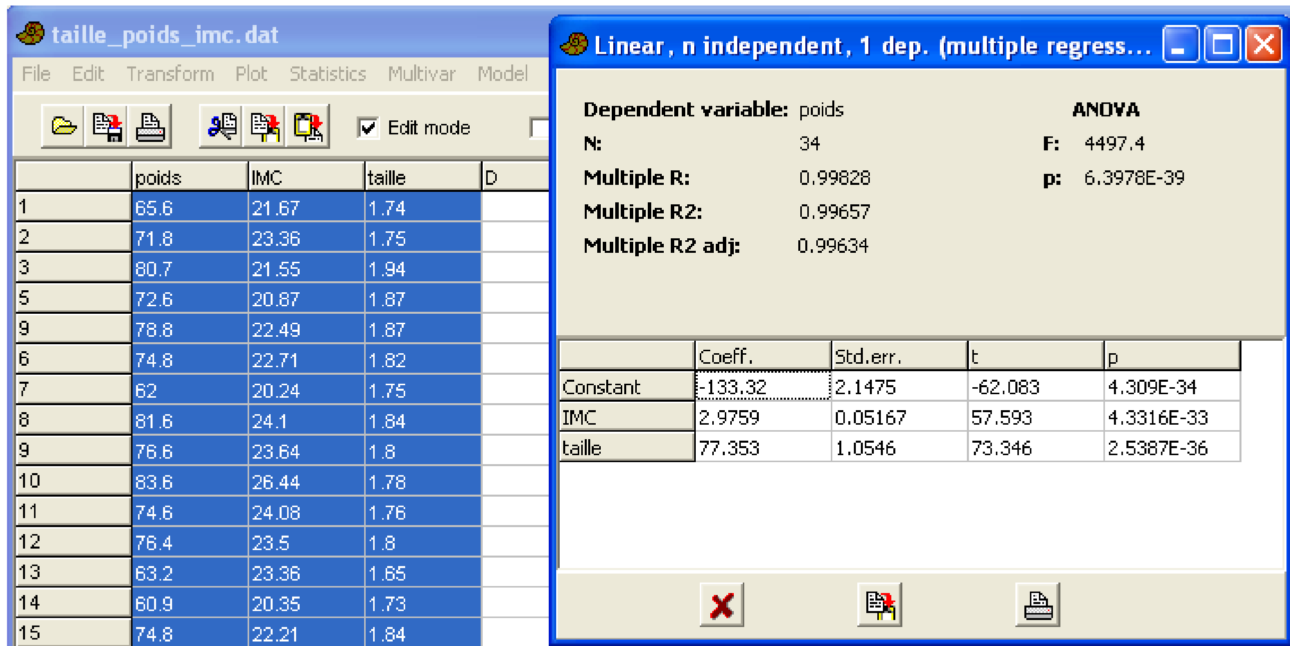


C'est très semblable à l'option précédente. La valeur indépendante (figurée par l'axe X) est la première colonne, et toutes les autres colonnes sont les variables dépendantes, figurées par l'axe Y. On peut passer de la figuration d'une variable dépendante à une autre par une paire de boutons, au dessus de « Overall MANOVA ».

Les valeurs de pente (Slope a) et d'ordonnée à l'origine (Interc b) sont calculées pour une régression standard (non RMA).

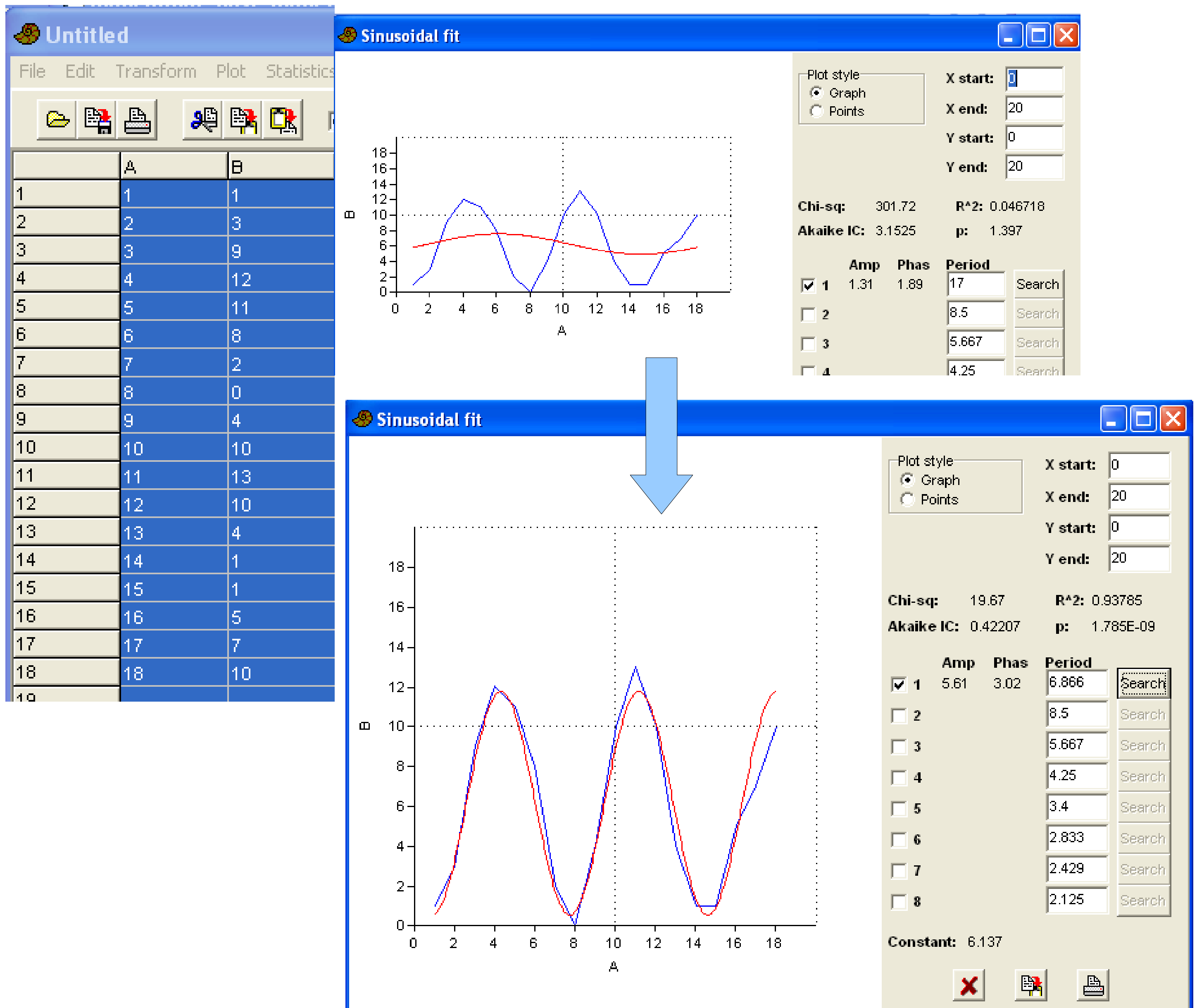
5.3 Linear n indep, 1 dep = régression linéaire multiple, pour expliquer une seule variable (dépendante) par n variables (indépendantes)

La variable dépendante doit être mise en première colonne, et toutes les autres variables (les autres colonnes) sont les variables explicatives (indépendantes).



Il n'y a pas de graphique, mais seulement un tableau donnant les valeurs numériques des coefficients à appliquer aux variables indépendantes pour prédire la variable dépendante, à ajouter à un terme constant (l'ordonnée à l'origine).

5.4 Sinusoidal = ajustement sinusoïdal de phénomènes périodiques



La fenêtre donne l'équation des sinusoïdes dont la somme aboutirait à expliquer les points mesurés.

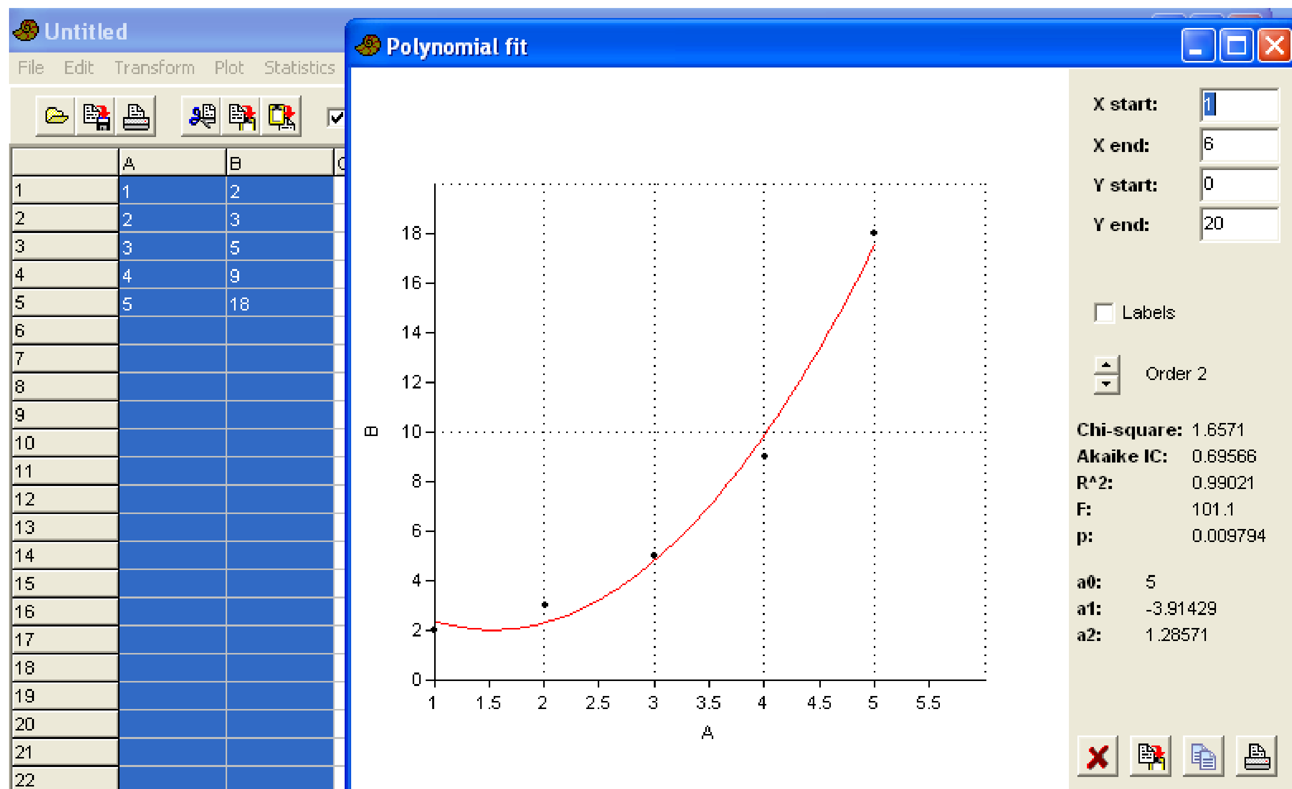
- en bas, la constante (valeur moyenne)
- vers le milieu, les diverses sinusoïdes dont la somme pourrait expliquer les points observés. Par défaut, seule la première est cochée, mais pour qu'elle soit bien ajustée, il faut cliquer sur « Search » pour que le logiciel cherche automatiquement la bonne fréquence. « Amp » indique l'amplitude (la moitié de l'intervalle entre minimum et maximum) et « Phas » indique la phase.

Pour une sinusoïde d'ordre 1, l'équation est de la forme
$y = C + a \cos(2\pi(x - x0)/T - \varphi)$
où C est la constante, a est l'amplitude, φ est la phase, T est la période, et x0 est la première valeur

de la série.

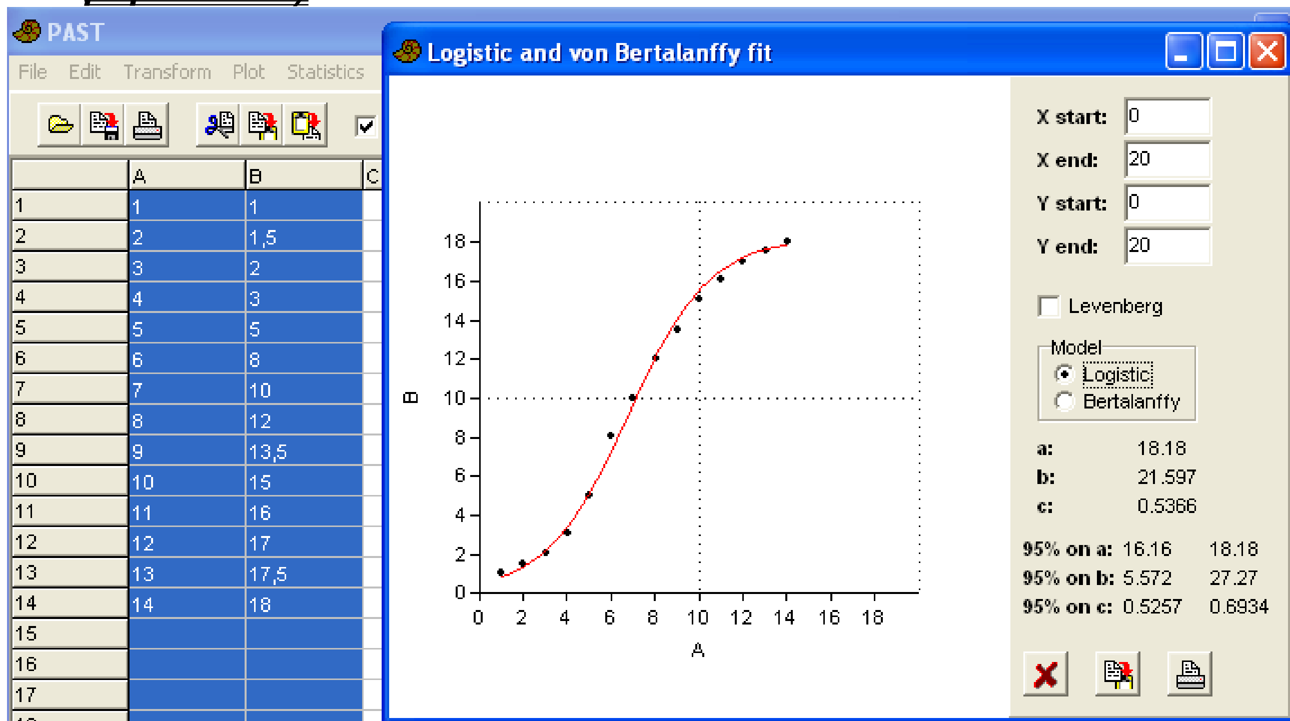
5.5 Polynomial = ajustement polynomial

L'équation du polynôme (par défaut de degré 2) est donnée par y = a0 + a1*x + a2*x².



On peut modifier la précision de l'ajustement (l'ordre du polynôme) par les deux boutons, au milieu de la zone de droite.

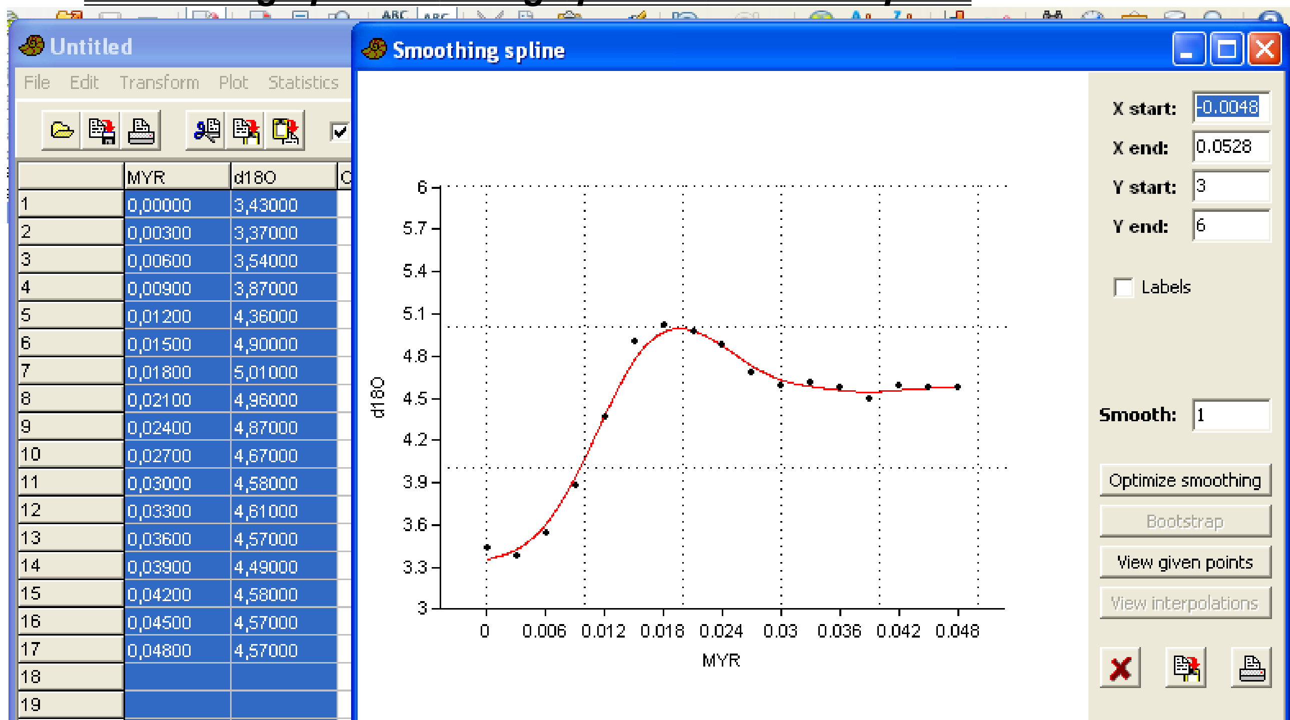
5.6 Logistic = ajustement à la courbe logistique (croissance de populations)



La courbe logistique est du type : $y = a/(1 + b\ e^{-cx})$.

En choisissant l'option « Bertalanffy », l'équation cherchée est du type : $y = a\ (1 - b\ e^{-cx})$.

5.7 Smoothing spline = lissage par les courbes spline



Ce n'est pas vraiment une méthode de modélisation, c'est une méthode de lissage, pour éliminer les bruits parasites des données brutes. Des points un peu trop dispersés sous l'effet de facteurs aléatoires soient résumés par une courbe plus jolie.

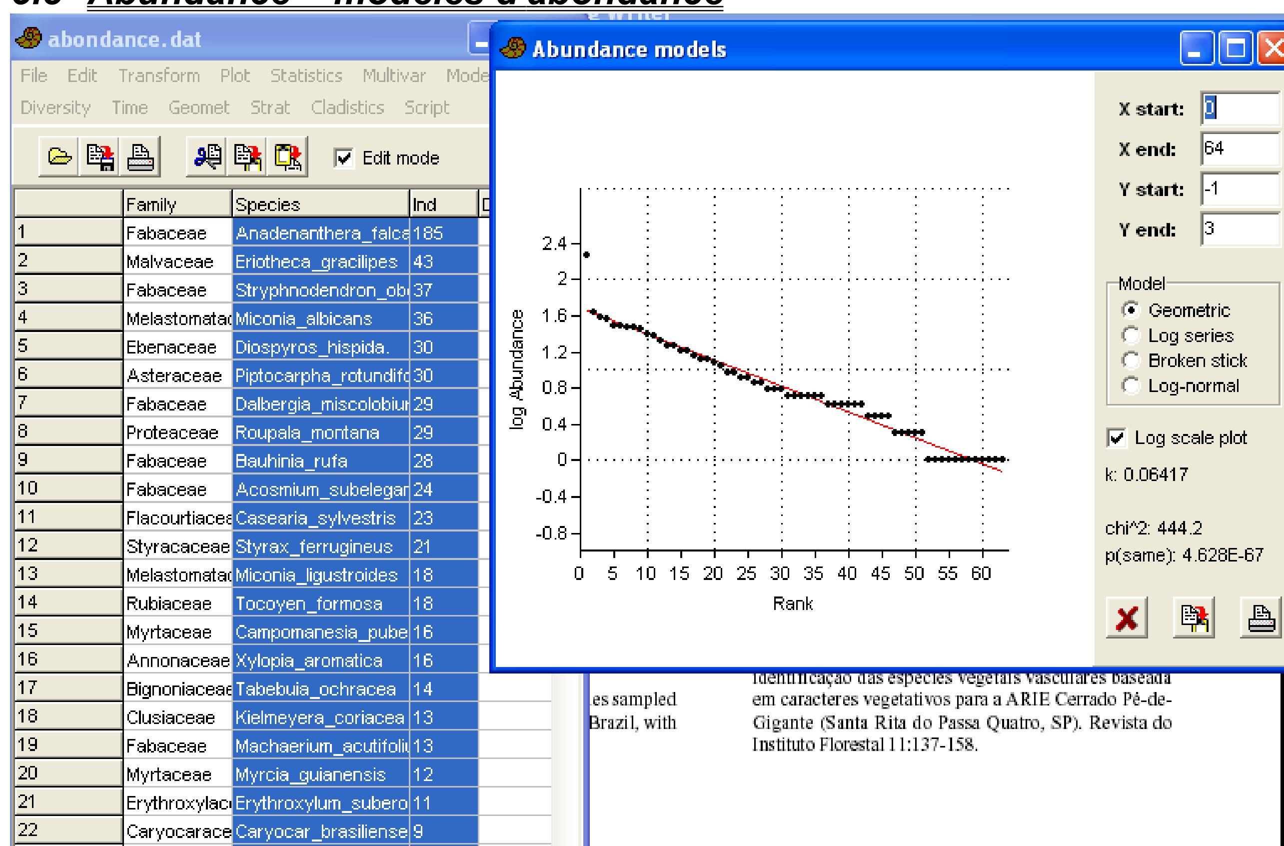
Le bouton « Optimize smoothing » oblige la courbe à passer par les divers points, mais le temps de calcul devient très long lorsqu'il y a beaucoup de points.

La valeur numérique saisie dans la ligne « Smooth: » indique la force du lissage. Par défaut elle est 1 (lissage assez faible), mais on peut y mettre des valeurs plus grandes (par exemple 5), ce qui aboutit à une courbe beaucoup plus lissée, s'écartant plus des points de mesure.

Le bouton « View given points » fait apparaître un tableau de valeurs numériques, avec les valeurs initiales de x et y, mais aussi les valeurs lissées des y. On peut sélectionner ces valeurs, les copier et les coller dans un tableur ou un autre tableau de PAST.

Il suffit de deux colonnes, x et y pour utiliser cette option. On peut aussi utiliser une troisième colonne, correspondant aux écarts-types des y,

5.8 Abundance = modèles d'abondance

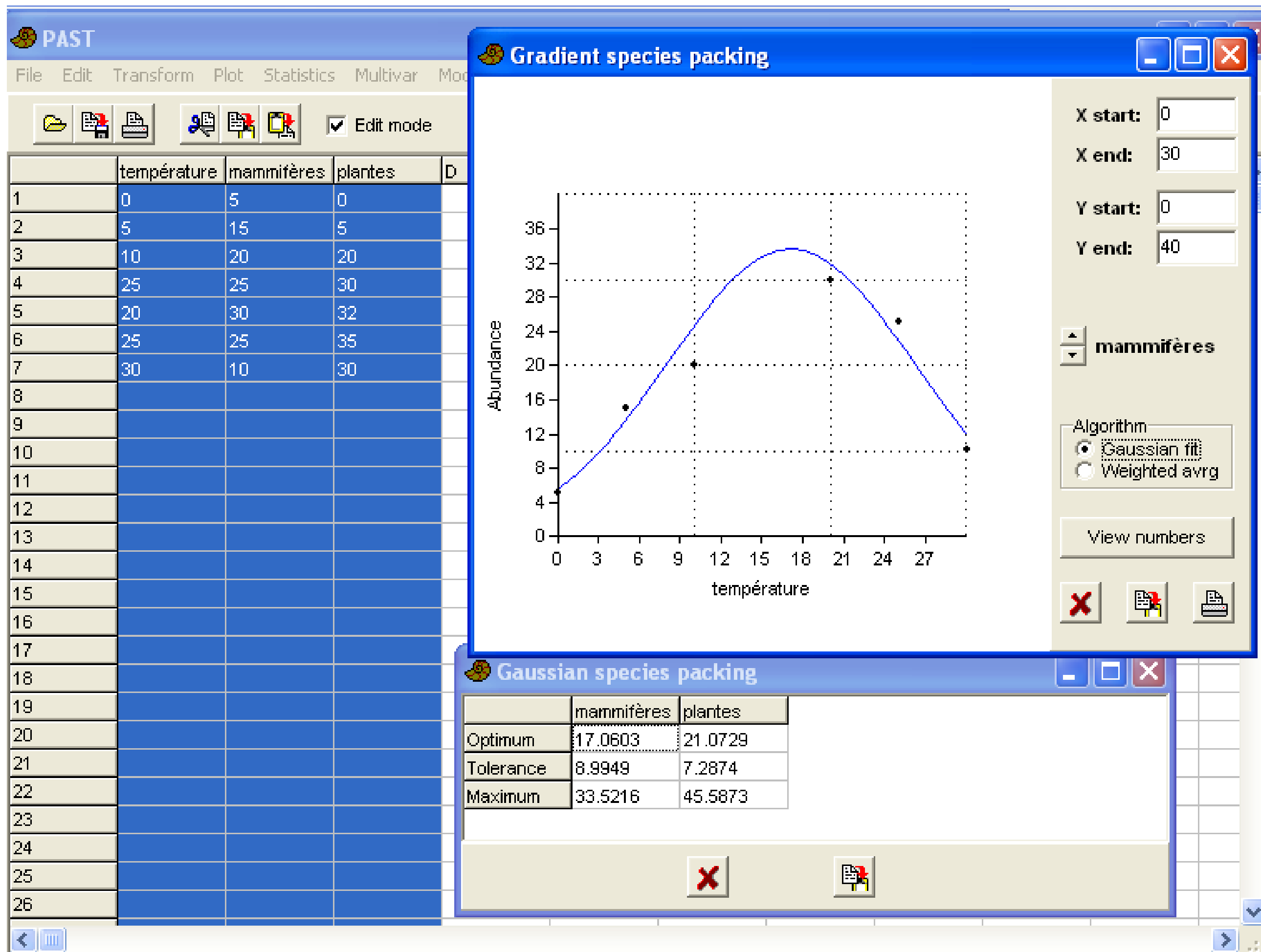


5.9 Species packing = « garniture d'espèces » : abondance d'espèces selon un gradient de facteur du milieu

Cette option permet un calcul théorique de l'optimum écologique pour un groupe systématique, ainsi que des intervalles de tolérance.

Deux modèles sont possibles pour le calcul de l'optimum :

- Par défaut, on suppose que la répartition écologique suit une loi normale en fonction du facteur du milieu (case « Gaussian » cochée)
- Un autre modèle possible est la moyenne pondérée (case « Weighted avrg » cochée).



	mammifères	plantes
Optimum	17.0603	21.0729
Tolerance	8.9949	7.2874
Maximum	33.5216	45.5873

	température	mammifères	plantes
1	0	5	0
2	5	15	5
3	10	20	20
4	25	25	30
5	20	30	32
6	25	25	35
7	30	10	30

6 Diversity : étude de la biodiversité

Diversity | Time | Geomet | St

- Diversity indices
- Quadrat richness
- Beta diversity
- Taxonomic distinctness
- Individual rarefaction
- Sample rarefaction
- Compare diversities
- Diversity t test
- Diversity profiles

Il existe diverses méthodes pour calculer la « biodiversité ». Le plus simple est la « richesse spécifique », qui est le nombre d'espèces présente sur une surface donnée. Par exemple, la richesse spécifique des Oiseaux terrestres en Amérique du Nord augmente du Canada vers l'isthme de Panama.

D'autres indices plus compliqués sont calculables.

6.1 Diversity indices = indices de diversité de type α (alpha) = biodiversité locale

C'est le plus simple à comprendre : pour un écosystème donné, on prend une surface fixe (un mètre carré pour les petits arthropodes du sol, un kilomètre carré pour les oiseaux, etc), et on compte le nombre d'individus d'espèces différentes sur cette surface.

Untitled

File Edit Transform Plot Statistics Multivar Model

Edit mode

	lieu1	lieu2	lieu3	lieu4
espèce_A	3	25	5	0
espèce_B	10	10	8	0
espèce_C	2	9	25	0
espèce_D	0	0	44	15
espèce_E	1	0	3	25
6				
7				
8				
9				
10				
11				
12				

Diversity

	lieu1	lieu2	lieu3	lieu4
Taxa_S	4	3	5	2
Individuals	16	44	85	40
Dominance_D	0.4453	0.4163	0.368	0.5313
Shannon_H	1.041	0.9825	1.208	0.6616
Simpson_1-D	0.5547	0.5837	0.632	0.4688
Evenness_e^H/S	0.7079	0.8904	0.6693	0.9689
Menhinick	1	0.4523	0.5423	0.3162
Margalef	1.082	0.5285	0.9004	0.2711
Equitability_J	0.7508	0.8943	0.7505	0.9544
Fisher_alpha	1.712	0.7286	1.161	0.4431
Berger-Parker	0.625	0.5682	0.5176	0.625

Bootstrap (95% confidence)

Cliquer la case « Bootstrap » augmente le nombre de colonnes : pour chaque lieu sont indiqués les limites supérieure et inférieure de l'intervalle de confiance pour les différents indices.

6.2 Quadrat richness = richesse des quadrats = richesse intralieu

Un quadrat est une parcelle d'échantillonnage écologique, en principe de forme carrée, d'où son nom. Selon le niveau de l'étude, la dimension d'un quadrat est variable : 100x100 m, ou 2x2 m, ou 1x1 m, etc. On peut par exemple utiliser des quadrats fixes, dont on suit la composition biologique au cours du temps, d'année en année.

	lieu1	lieu2	lieu3	lieu4
espèce_A	1	1	0	1
espèce_B	1	1	1	1
espèce_C	1	1	1	1
espèce_D	0	1	1	1
espèce_E	1	1	1	1

Quadrat species richness estimators

Chao 2: 5 Chao 2 variance: 0
Jackknife 1: 5
Jackknife 2: 5
Bootstrap: 5.00781

On réalise donc un tableau de présence/absence des différentes espèces (en lignes) dans les différents quadrats (les colonnes), et divers indices sont calculés. Plus les quadrats sont différents les uns des autres, plus ces indices sont grands.

6.3 Beta diversity =indices de diversité de type beta = diversité interlieux

	lieu1	lieu2	lieu3	lieu4
espèce_A	3	25	5	0
espèce_B	10	10	8	0
espèce_C	2	9	25	0
espèce_D	0	0	44	15
espèce_E	1	0	3	25

Beta diversity

Whittaker: 0.42857
Harrison: 0.10714
Cody: 2
Routledge: 0.12727
Wilson-Shmida: 0.71429
Mourelle: 0.17857
Harrison 2: 0.083333
Williams: 0.25

Alors que la diversité de type α était mesurée lieu par lieu, la diversité de type β mesure la variabilité entre divers lieux selon plusieurs méthodes.

6.4 Taxonomic distinctness = distance taxonomique

	espèce	lieu1	lieu2	lieu3	lieu4
1	espèce_A	3	25	0	8
2	espèce_B	10	0	8	9
3	espèce_C	0	0	25	10
4	espèce_D	0	0	44	15
5	espèce_E	0	0	3	25

Taxonomic distinctness with 95% confidence intervals

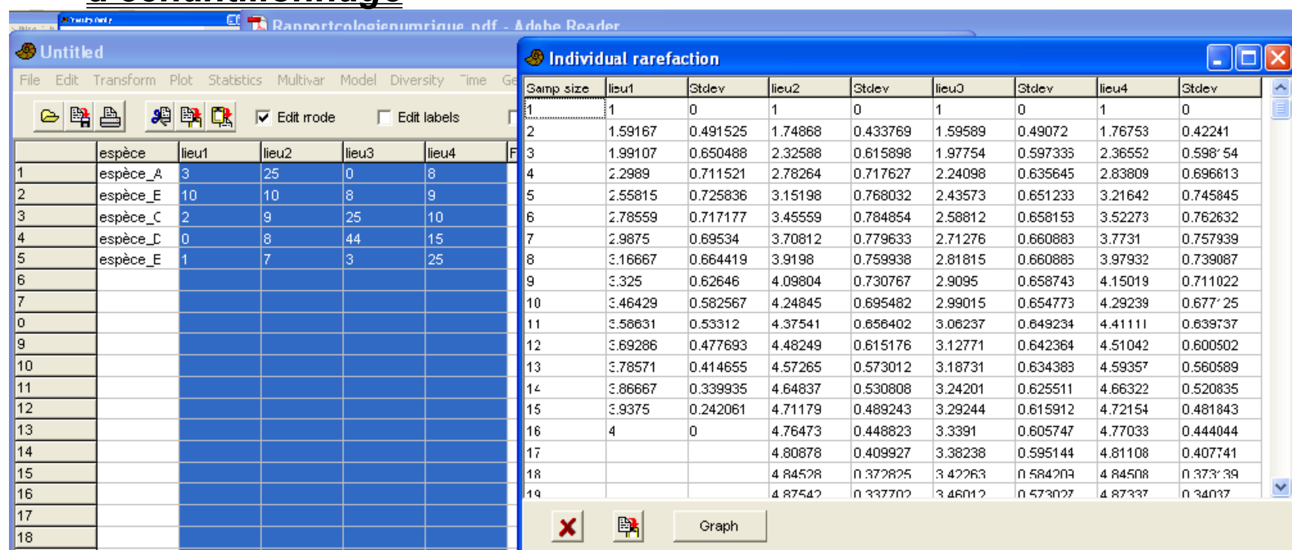
	lieu1	lieu2	lieu3	lieu4
Diversity	0.7692		1.192	1.535
Lower limit	1.231		1.499	1.492
Upper limit	1.718		1.608	1.613
Distinctness	2		2	2
Lower limit	2		2	2
Upper limit	2		2	2

Ces indices donnent des propriétés des différents sites de mesure, et sont fondés sur la classification linnéenne.

Il faut mettre dans la colonne de gauche les noms d'espèce. Éventuellement, dans la colonne suivante on peut mettre les noms de genre (de ces espèces), puis dans la colonne suivante les noms de famille, etc. Toutes les autres colonnes correspondent à des lieux, et on met dans les cases les comptages des individus des différents groupes systématiques.

« diversity » : c'est une sorte de moyenne des longueurs des chemins entre les paires d'individus de l'échantillon (de taxons différents)
« distinctness » : c'est une sorte de moyenne des longueurs des chemins entre les paires d'individus de l'échantillon (y compris du même taxon).

6.5 *Individual rarefaction : estimation de l'effet de l'effort d'échantillonnage*



	espèce	lieu1	lieu2	lieu3	lieu4
1	espèce_A	3	25	0	8
2	espèce_B	10	10	8	9
3	espèce_C	2	9	25	10
4	espèce_D	0	8	44	15
5	espèce_E	1	7	3	25

Samp size	lieu1	Stdev	lieu2	Stdev	lieu3	Stdev	lieu4	Stdev
1	1	0	1	0	1	0	1	0
2	1.59167	0.491525	1.74868	0.433769	1.59589	0.49072	1.76753	0.42241
3	1.99107	0.650488	2.32588	0.615898	1.97754	0.597333	2.36552	0.598154
4	2.2989	0.711521	2.78264	0.717627	2.24098	0.635645	2.83809	0.696613
5	2.55815	0.725836	3.15198	0.768032	2.43573	0.651233	3.21642	0.745845
6	2.78559	0.717177	3.45559	0.784854	2.58812	0.658153	3.52273	0.762632
7	2.9875	0.69534	3.70812	0.779633	2.71276	0.660883	3.7731	0.757939
8	3.16667	0.664419	3.9198	0.759938	2.81815	0.660883	3.97932	0.739087
9	3.325	0.62646	4.09804	0.730767	2.9095	0.658743	4.15019	0.711022
10	3.46429	0.582567	4.24845	0.695482	2.99015	0.654773	4.29239	0.677125
11	3.58631	0.53312	4.37541	0.656402	3.06237	0.649234	4.41111	0.639737
12	3.69286	0.477693	4.48249	0.615176	3.12771	0.642364	4.51042	0.600502
13	3.78571	0.414655	4.57265	0.573012	3.18731	0.634383	4.59357	0.560589
14	3.86667	0.339935	4.64837	0.530808	3.24201	0.625511	4.66322	0.520835
15	3.9375	0.242061	4.71179	0.489243	3.29244	0.615912	4.72154	0.481843
16	4	0	4.76473	0.448823	3.3391	0.605747	4.77033	0.444044
17			4.80878	0.409927	3.38238	0.595144	4.81108	0.407741
18			4.84528	0.372825	3.42263	0.584209	4.84508	0.373139
19			4.87542	0.337702	3.46012	0.573027	4.87337	0.34037

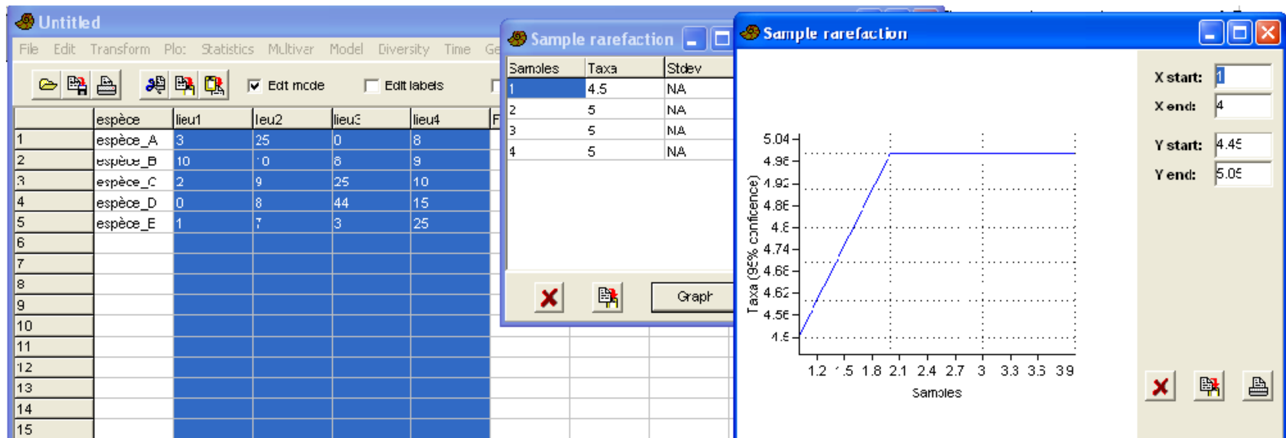
Cette fonction permet d'imaginer quelle aurait été la richesse spécifique d'une récolte (le nombre d'espèces espéré) si l'effort d'échantillonnage avait été moins important.

« Samp size » donne la taille de l'échantillon que l'on aurait pu réaliser :

- avec 1 seul individu capturé, bien sûr, la richesse spécifique vaut 1, et son écart-type vaut 0.
- avec 2 ou plusieurs individus capturés, la richesse spécifique attendue augmente avec le nombre d'individus capturés, de façon variable selon les populations.
- Le maximum calculable correspond au nombre d'individus dans la population initialement mesurée.

6.6 *Sample rarefaction (courbe d'accumulation spécifique) : indice de Mao-tau*

Là encore, on a divers lieux d'observation en colonnes, et divers groupes taxonomiques en lignes.



Cette fonction donne la richesse spécifique en fonction du nombre d'échantillons testés.

6.7 Compare diversities : comparaison des diversités

espèce	lieu1	lieu2
espèce_A	3	25
espèce_B	10	10
espèce_C	2	9
espèce_D	0	8
espèce_E	1	7

Compare diversities

	lieu1	lieu2	Boot p(eq)	Perm p(eq)
Taxa S	4	5	0.385	0.309
Individuals	16	59	0	0
Dominance	0.4453	0.264	0.035	0.034
Shannon H	1.041	1.475	0.047	0.042
Evenness e^H/S	0.7079	0.8745	0.1	0.087
Simpson indx	0.5547	0.736	0.035	0.034
Menhinick	1	0.6509	0.619	0.691
Margalef	1.082	0.981	1	1
Equitability J	0.7508	0.9167	0.054	0.037
Fisher alpha	1.712	1.304	0.961	0.971
Berger-Parker	0.625	0.4237	0.092	0.069

A partir des indices de diversité alpha, cette option calcule les probabilités que les deux lieux soient identiques (dans les deux colonnes de gauche, par deux méthodes différentes).

6.8 Diversity t test : test t de diversité

espèce	lieu1	lieu2
espèce_A	3	25
espèce_B	10	10
espèce_C	2	9
espèce_D	0	8
espèce_E	1	7

Shannon diversity t test

	lieu1	lieu2
S:	4	5
Index:	0.94709	1.4414
Variance:	0.043428	0.0054818

t: -2.2353 p(same): 0.036838

Cette option calcule la probabilité que les indices de diversité de Shannon soient les mêmes pour les

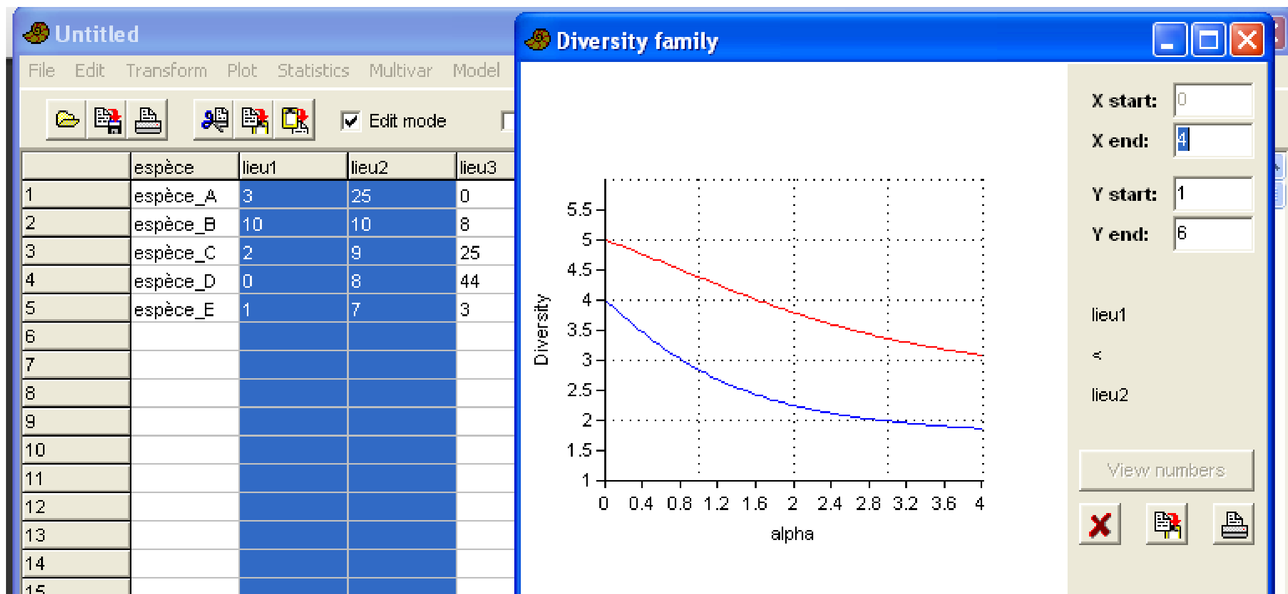
deux échantillons.

6.9 Diversity profiles : profils de diversité

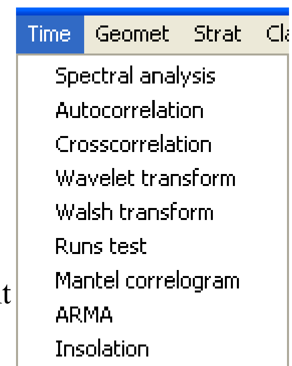
Il peut exister diverses manières de calculer un indice de biodiversité, et l'indice de Shannon n'en est qu'une d'entre elles.

Cette option calcule la biodiversité selon plusieurs façons, et affiche le résultat sous forme d'une courbe.

- pour α = 0, les indices sont simplement le nombre d'espèces
- pour α = 1, les indices sont proportionnels aux indices de Shannon
- pour α =2, les indices se comportent comme les indices de Simpson

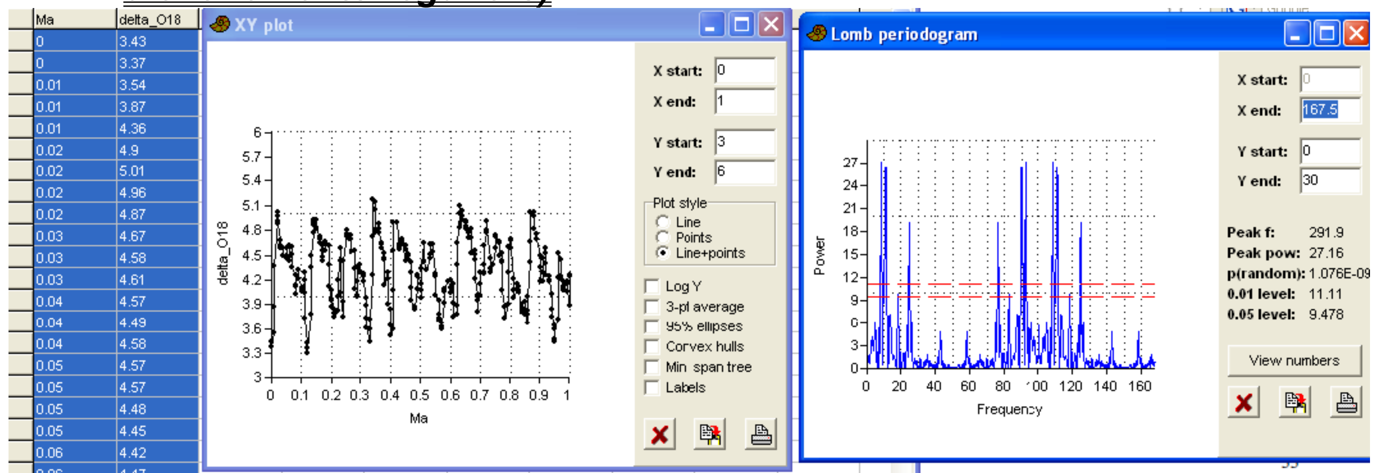


7 Time : étude des séries temporelles



Les séries temporelles, aussi appelées séries chronologiques, correspondent à des valeurs numériques mesurées en fonction du temps : il y a obligatoirement un classement, qui est le déroulement du temps. Selon les cas, le temps peut être mesuré en microsecondes ou en millions d'années. La spécificité de l'analyse des séries temporelles est de pouvoir mettre en évidence des répétitions et des régularités au cours du temps.

7.1 Spectral analysis = analyse spectrale (calcul des fréquences d'événements réguliers)



Les données doivent être dans deux colonnes : la première est le temps, la seconde est les valeurs observées.

Cette option réalise une transformée de Fourier pour calculer l'importance des fréquences des événements observés. Elle trace en abscisses la fréquence (nombre de cycles par unité de temps), et en ordonnées la puissance des cycles en question. L'exemple ci-dessus correspond à la concentration d'oxygène 18, qui indique des changements climatiques au cours du dernier million d'années. Les pics correspondent aux cycles de Milankovitch.

Il faut au moins quatre cycles pour que cette fonction soit correcte.

- S'il n'y a qu'une seule colonne de mesure, il faut que les mesures soient régulièrment espacées. L'estimation du temps est faite par les numéros des lignes.
- S'il y a deux colonnes (le temps dans la première, et les valeurs observées dans la seconde), l'algorithme utilisé permet de travailler même sur des données où les intervalles de temps entre les mesures sont irréguliers, ce qui est fréquent dans les sciences de la nature, et en particulier en paléontologie.

7.2 Autocorrelation = autocorrélation : les mesures sont elles liées au cours du temps ?

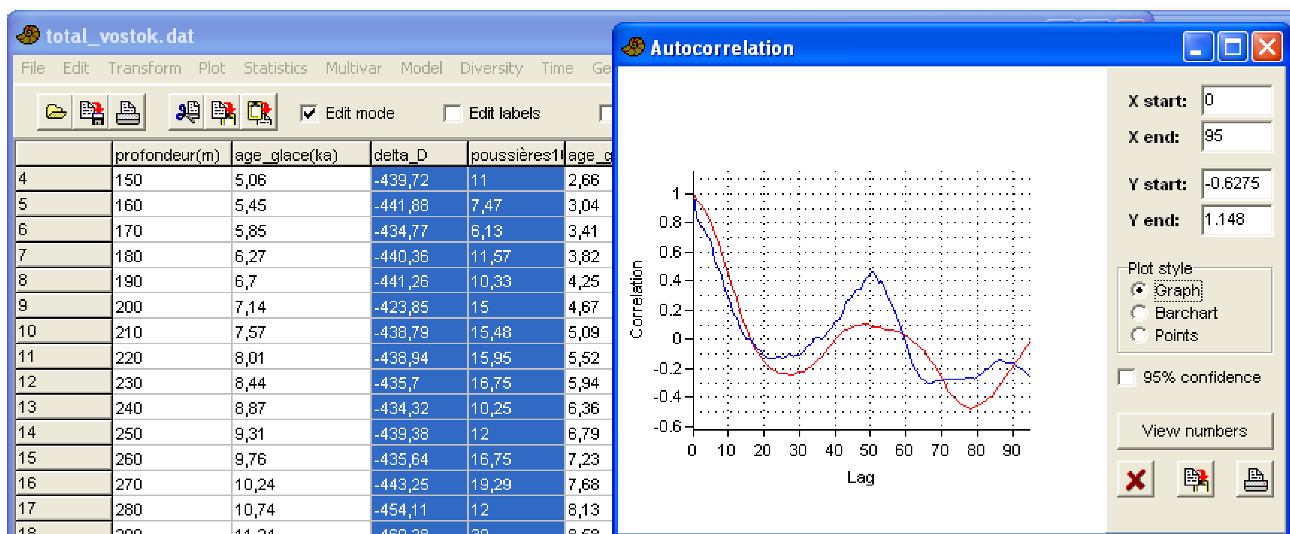
L'étude de l'autocorrélation, découlant de l'autocovariance, essaie de répondre à la question : la mesure à un instant est-elle liée à la mesure faite à l'instant suivant ou à l'instant précédent ?

- Dans certains cas, il n'y a aucune corrélation au cours du temps, à aucune échelle, par exemple lors d'un lancer de dé.
- Dans d'autres cas, il y a une forte corrélation pour les temps proches, mais une faible corrélation pour les temps longs. Il peut même exister une corrélation négative pour certains intervalles de temps, s'il existe des phénomènes périodiques. Par exemple, à un intervalle d'une journée, il existe une corrélation positive pour les températures mesurées à l'extérieur des maisons, mais à un intervalle de six mois, il y a une corrélation négative puisque l'été est six mois après l'hiver.

Ainsi, sous une autre forme, cette étude de l'autocorrélation donne des renseignements du même type que l'analyse spectrale.

Il suffit de deux cycles pour permettre l'étude de l'autocorrélation, mais il faut que les mesures soient régulièrement espacées.

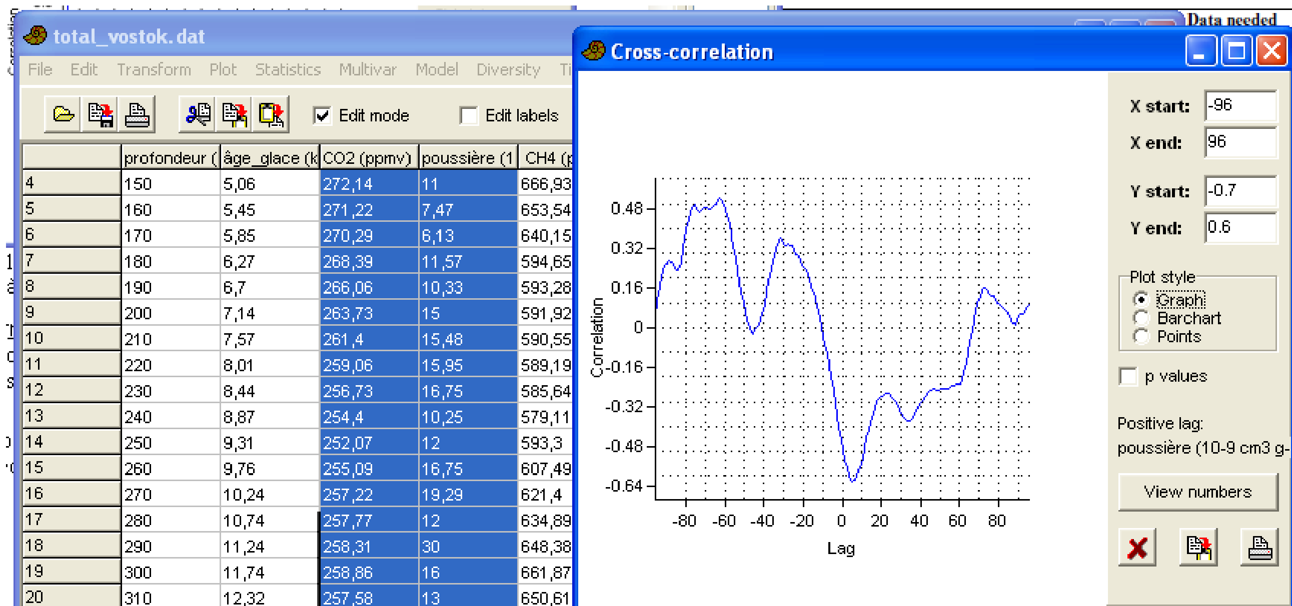
L'abscisse (« Lag ») correspond au numéro des mesures, et va jusqu'à la moitié du nombre de mesures traitées.



Il suffit d'une seule colonne de valeurs mesurées, puisque les abscisses du graphique correspondent aux numéros des mesures, et non à une durée temporelle.

7.3 Crosscorrelation = corrélation croisée entre deux variables

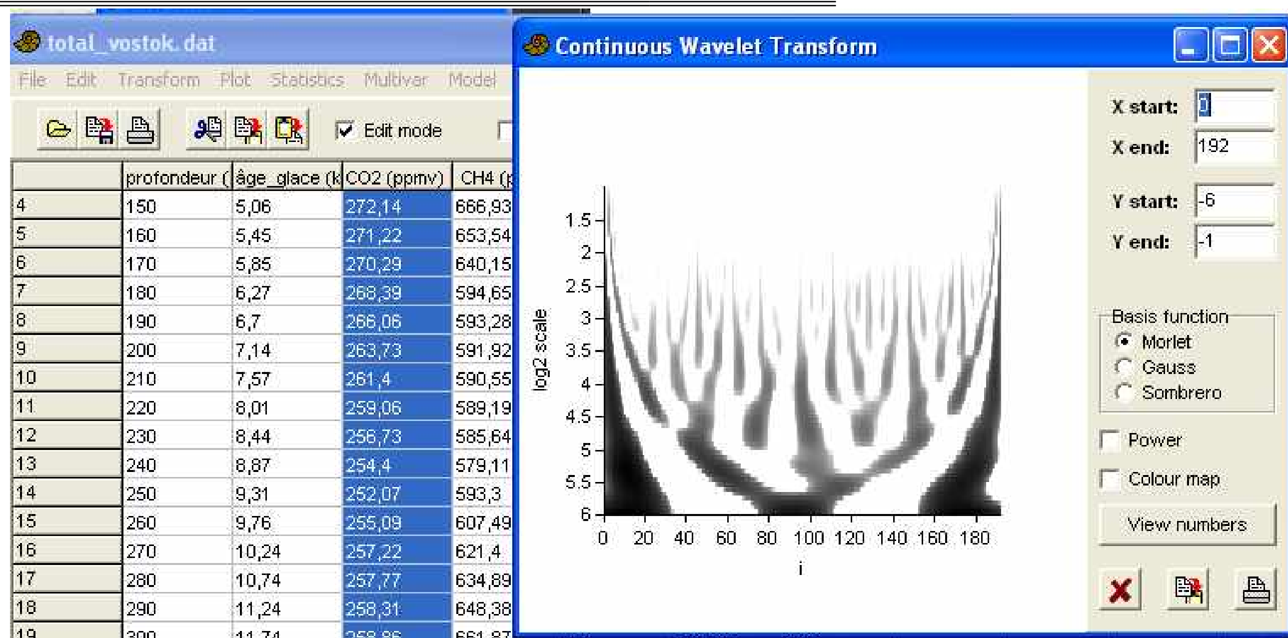
Il faut fournir deux colonnes de données mesurées régulièrement, et cette fonction cherche à aligner les deux colonnes de données, c'est à dire à trouver le décalage optimal pour que les deux colonnes soient bien corrélées.



	profondeur (	âge_glace (k	CO2 (ppmv)	poussière (1	CH4 (p
4	150	5,06	272,14	11	666,93
5	160	5,45	271,22	7,47	653,54
6	170	5,85	270,29	6,13	640,15
7	180	6,27	268,39	11,57	594,65
8	190	6,7	266,06	10,33	593,28
9	200	7,14	263,73	15	591,92
10	210	7,57	261,4	15,48	590,55
11	220	8,01	259,06	15,95	589,19
12	230	8,44	256,73	16,75	585,64
13	240	8,87	254,4	10,25	579,11
14	250	9,31	252,07	12	593,3
15	260	9,76	255,09	16,75	607,49
16	270	10,24	257,22	19,29	621,4
17	280	10,74	257,77	12	634,89
18	290	11,24	258,31	30	648,38
19	300	11,74	258,86	16	661,87
20	310	12,32	257,58	13	650,61

Le graphique obtenu montre en abscisses le décalage (retard) entre les deux séries de mesures, et en ordonnées la corrélation entre les deux séries pour ce décalage.

7.4 Wavelet transform : transformation en ondelettes

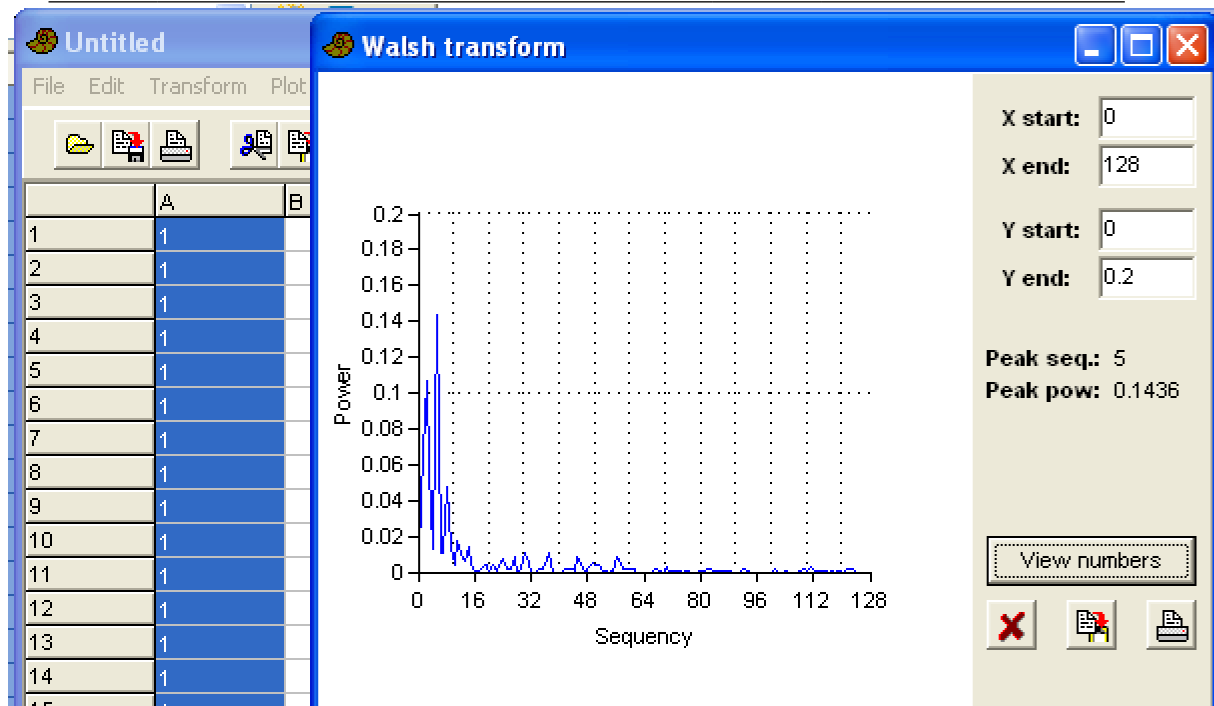


	profondeur (	âge_glace (k	CO2 (ppmv)	CH4 (p
4	150	5,06	272,14	666,93
5	160	5,45	271,22	653,54
6	170	5,85	270,29	640,15
7	180	6,27	268,39	594,65
8	190	6,7	266,06	593,28
9	200	7,14	263,73	591,92
10	210	7,57	261,4	590,55
11	220	8,01	259,06	589,19
12	230	8,44	256,73	585,64
13	240	8,87	254,4	579,11
14	250	9,31	252,07	593,3
15	260	9,76	255,09	607,49
16	270	10,24	257,22	621,4
17	280	10,74	257,77	634,89
18	290	11,24	258,31	648,38
19	300	11,74	258,86	661,87

C'est une méthode qui permet de chercher des régularités (périodicités) à la fois à différentes échelles de fréquences, pour une seule colonne de données, supposées régulièrement réparties.

En abscisses, les numéros des mesures (donc le temps, s'il s'agit vraiment de mesures chronologiques). En ordonnée, la période, selon une échelle logarithmique. Les taches foncées indiquent une forte variabilité pour la période considérée, et le fond clair indique une faible variabilité, en bas avec une vue détaillée, en haut avec une vue globale.

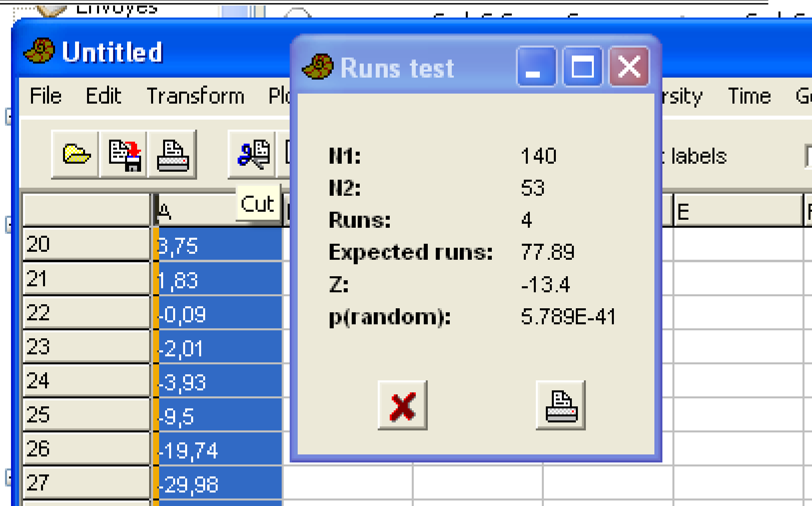
7.5 Walsh transform = transformation de Walsh ou de Hadamard



C'est une sorte d'analyse spectrale par transformée de Fourier pour des données binaires ou ordinales. Il faut utiliser une seule colonne, de données régulièrement réparties.

C'est une fonction à employer avec précautions, car les résultats ne sont pas très interprétables.

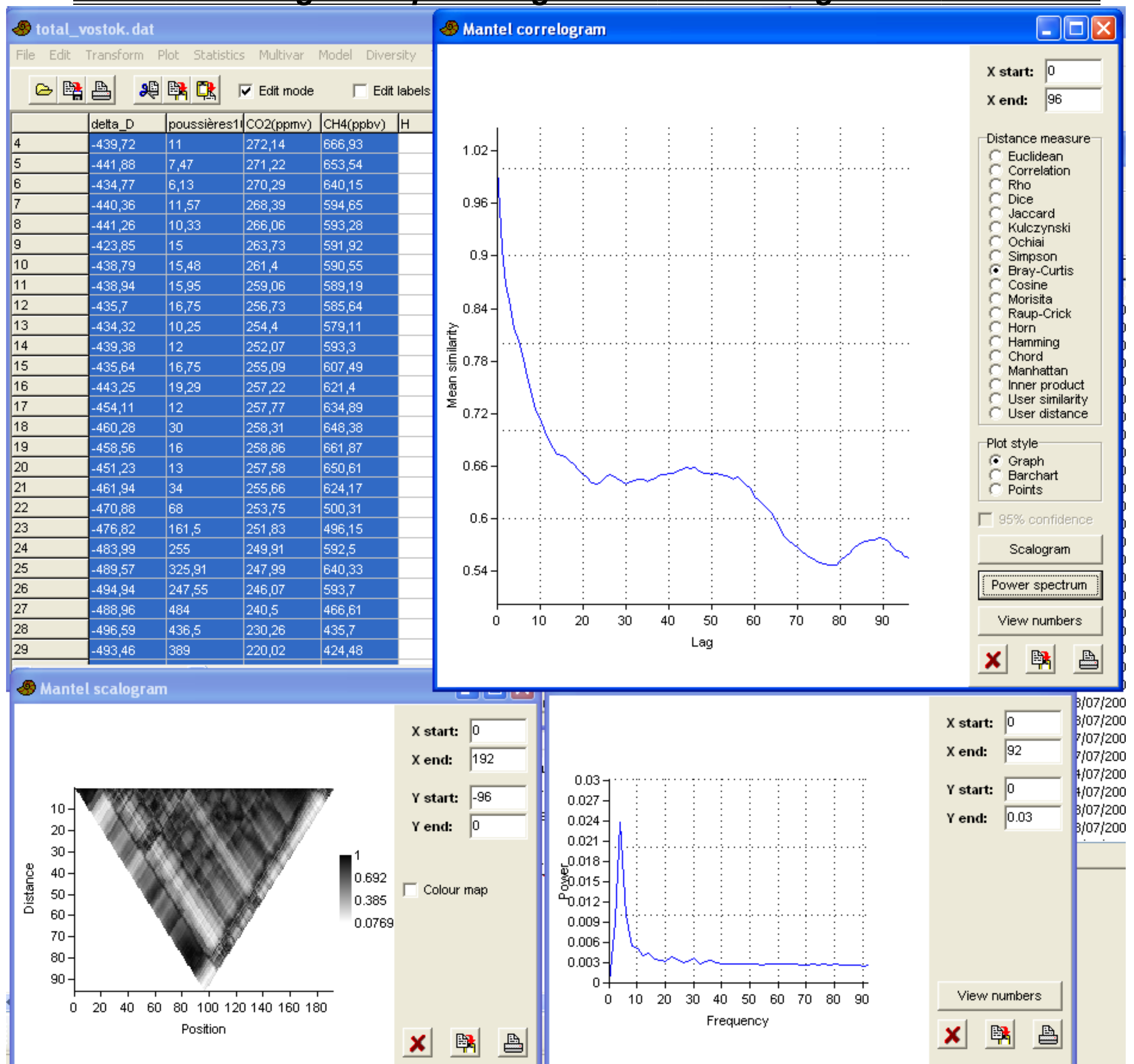
7.6 Runs test = test d'aléatoirité d'une série de données



C'est un test non paramétrique qui teste si une série est au hasard ou non. « non au hasard » peut être aussi bien une autocorrélation, une tendance ou une périodicité.

Il faut lui fournir une série de données, positives et négatives. Les valeurs positives sont converties en 1, et les valeurs négatives ou nulles en 0.

7.7 Mantel correlogram = périodogramme et corrélogramme de Mantel



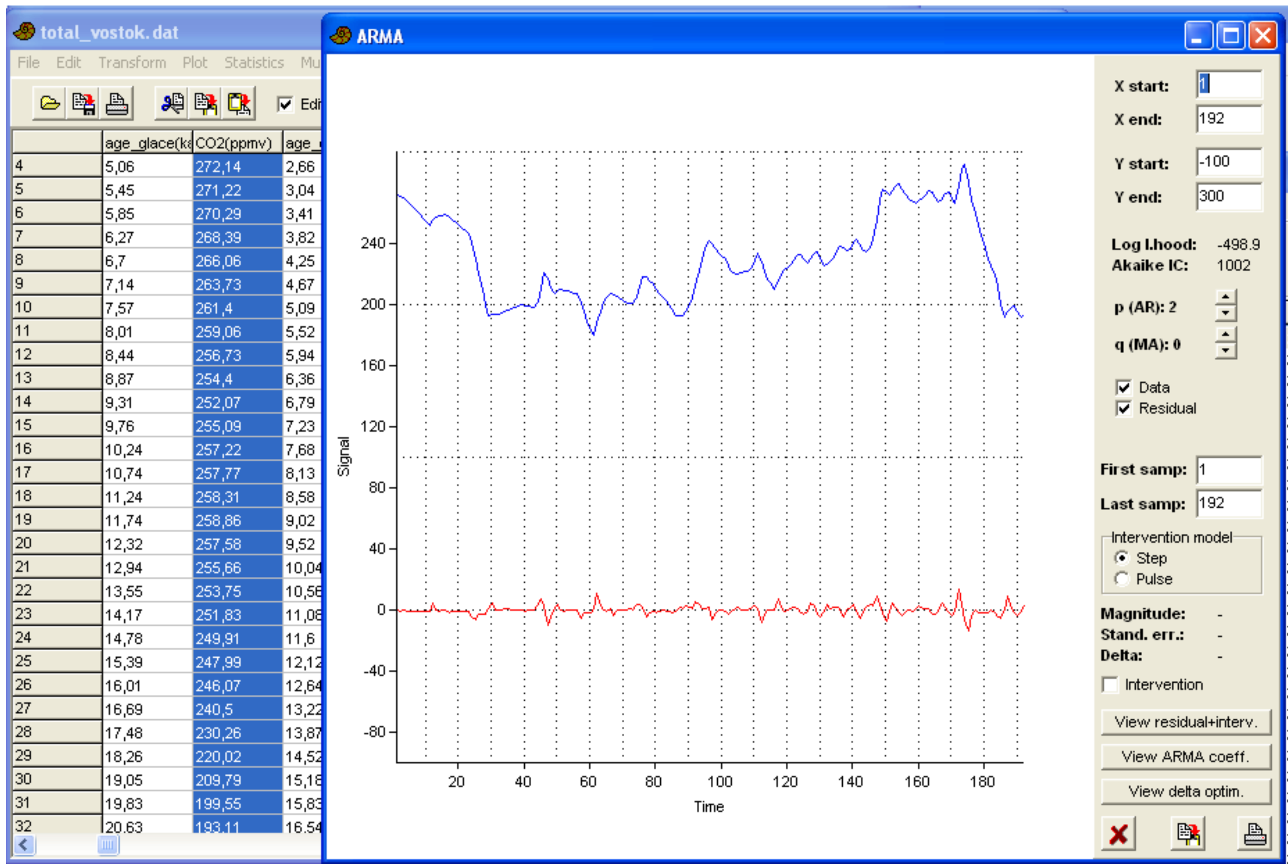
C'est une extension multivariée à l'autocorrélation, basée sur des similarités et mesures de distances (nombreux types disponibles).

Le corrélogramme montre la similarité moyenne entre les séries temporelles et une copie décalée dans le temps, pour différents décalages.

7.8 ARMA

ARMA est la combinaison de « AR » (autorégression) et « MA » (mobile average = moyenne mobile). C'est un module puissant, mais assez compliqué, à partir d'une colonne de données régulièrement espacées.

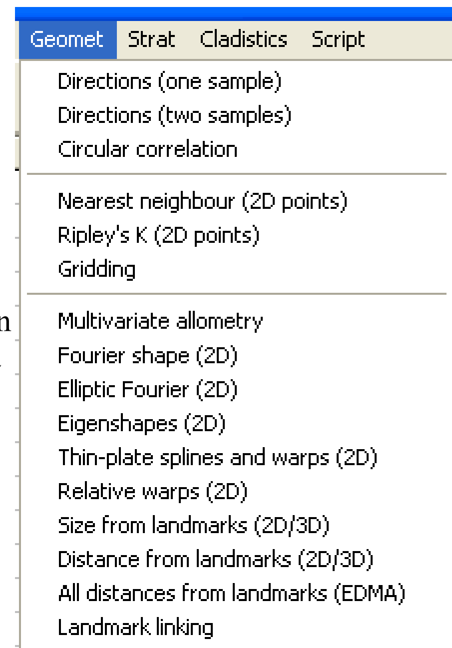
Deux courbes sont affichées : en bleu le simple graphique des données, en rouge les résidus, en fonction des numéros des données.



7.9 *(Insolation)*

(fonction spéciale pour calculer l'insolation d'une région à une époque donnée de l'histoire de la Terre ; il faut des données supplémentaires : voir fichier d'aide de PAST)

8 Geomet = mesures géométriques

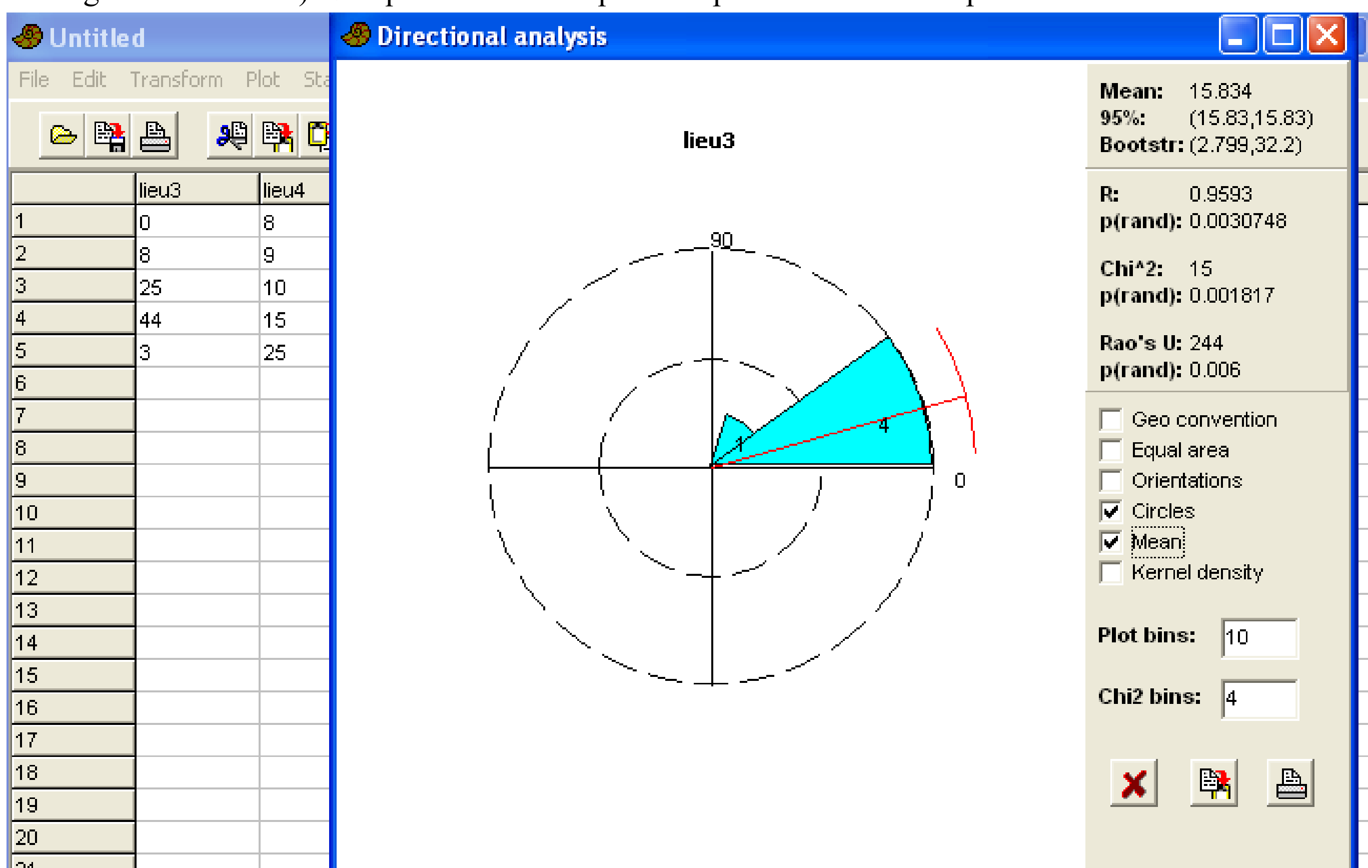


Ce chapitre est destiné à répondre à des questions du type : « les mesures d'angle sont-elles concentrées, ou bien sont-elles aléatoires, dans n'importe quelle direction de l'espace ? », ou bien « à partir d'une dizaine de points de mesures sur l'ensemble de la carte, peut-on établir une carte de pollution ? ».

8.1 Directions (one sample) = étude des orientations pour un échantillon

La colonne étudiée doit contenir les mesures d'orientation, en degrés.

Cette option compte le nombre de mesures faisant partie d'un secteur (par défaut 10 secteurs, mais on peut fixer le nombre par la ligne « Plot bins »). Chaque secteur est représenté par une taille correspondant au nombre de



	lieu3	lieu4
1	0	8
2	8	9
3	25	10
4	44	15
5	3	25

mesures qui y ont été observées.

Par défaut, l'orientation est avec la convention mathématique : le zéro est à droite (vers l'est), et le sens est antihoraire. On peut choisir la convention géographique, zéro vers le haut (vers le nord) et sens horaire en cochant la case « Geo convention ».

Par défaut, le nombre d'observations dans un secteur est rendu par la longueur du rayon de ce

secteur. Si on veut qu'il soit rendu par la surface du secteur, cocher la case « Equal area ».

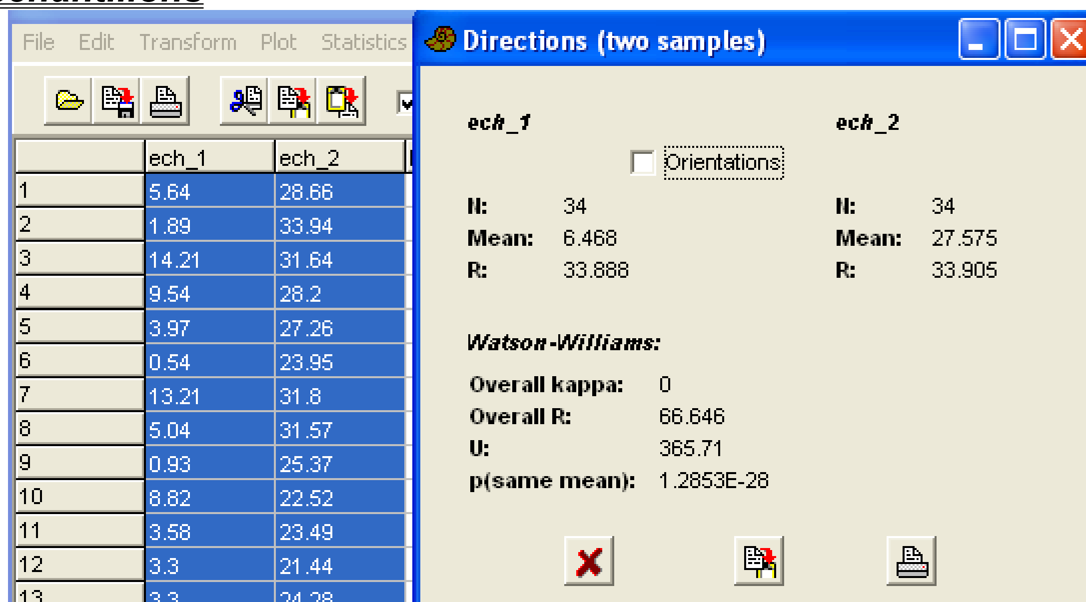
Si on coche la case « orientations », les secteurs symétriques sont aussi tracés.

La case « Circles » permet de tracer les cercles de repérage, la case « Mean » permet de tracer la direction moyenne des observations.

Diverses valeurs statistiques sont dans la partie droite du graphique. En haut, la moyenne et l'intervalle de confiance à 95%, ainsi que l'intervalle de confiance par rééchantillonnage.

On voit ensuite trois façons différentes d'estimer la probabilité que la distribution soit due au hasard. L'estimation par la méthode du Khi-2 dépend de la valeur mise dans la case « Chi2 bins ».

8.2 Directions (two samples) = comparaison des directions de deux échantillons



Les directions de deux échantillons sont comparées, et la probabilité que ces deux directions soient identiques est affichée à p(same mean): Pour que ce test soit valable, il faut que les variances angulaires soient similaires.

Par défaut, on suppose que les valeurs sont des directions (valeurs possibles entre 0 et 360), mais en cochant la case « Orientations », on indique que les valeurs sont des orientations (valeurs possibles entre 0 et 180).

8.3 Circular correlation = corrélation angulaire

Contrairement au test précédent, il n'y a pas besoin de distributions semblables, mais il faut de nombreuses mesures pour qu'on puisse affirmer que les deux échantillons ne sont pas corrélés. Là encore, en cochant la case « Orientations », on indique que les valeurs sont des orientations (l'orientation Nord-Sud est équivalente à l'orientation Sud-Nord).

8.4 Nearest neighbour (2D points) = tests de la répartition aléatoire sur une surface

On doit fournir à cette fonction un couple de colonnes, qui correspond aux coordonnées bidimensionnelles x/y des points d'observation.

La fonction calcule les distances entre les divers points, et teste si la répartition est aléatoire ou non.

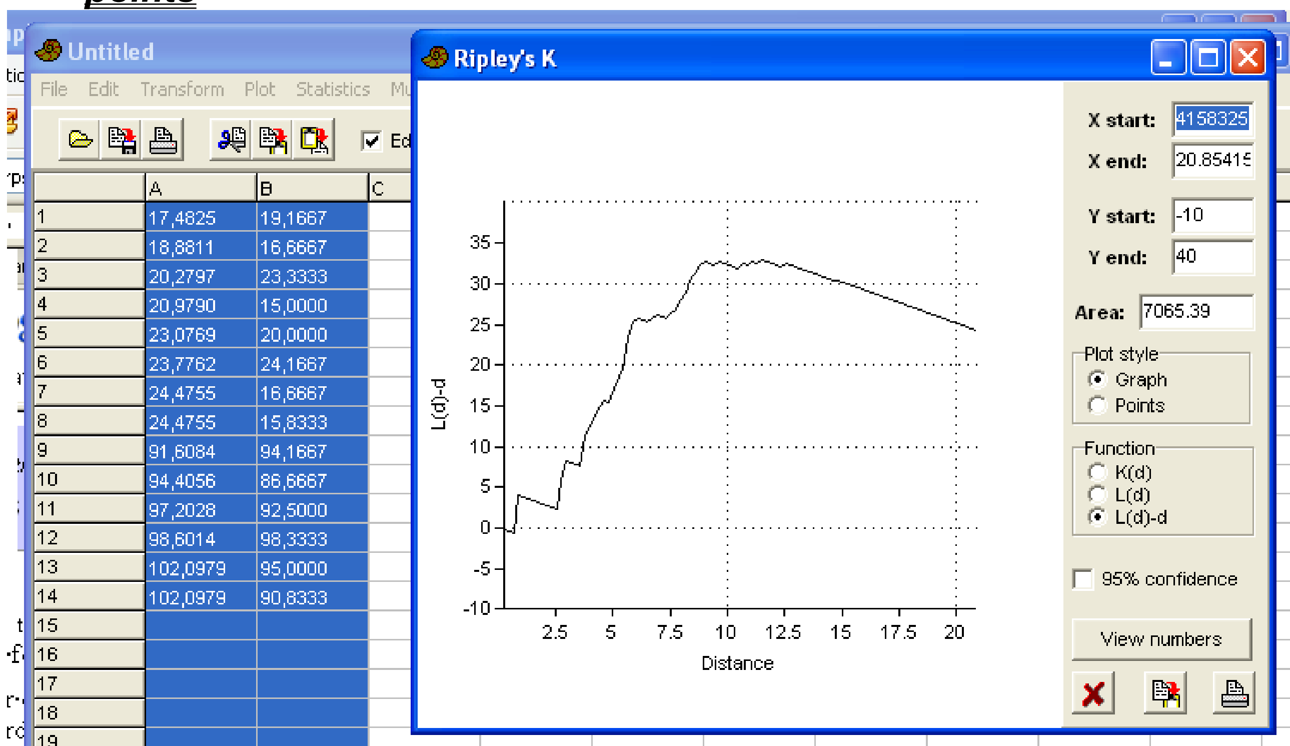
Par défaut, les calculs sont faits sur l'enveloppe convexe (« convex hull »). Il y a une correction de bordure (« Edge correction »), destinée à tenir compte du fait que les points situés à la périphérie ont forcément une distance plus grande aux autres points que les points situés dans la partie

centrale, mais on peut modifier cette option.

La distance moyenne de chaque point à son plus proche voisin est calculée (« Mean distance »), et comparée à une valeur théorique, ce qui permet de calculer la probabilité que la distribution soit aléatoire (« p(random) »).

Si les points sont groupés en agrégats, la valeur de R est inférieure à 1 ; si les point sont dispersés aléatoirement selon une loi de Poisson, R vaut 1, et si les points sont dispersés uniformément, R est supérieur à 1.

8.5 Ripley's K (2D points) = indication visuelle du groupement des points



Là aussi, on fournit à la fonction un couple de colonnes, correspondant aux coordonnées des points en x/y.

PAST calcule les distances entre les points, et trace en fonction de la distance une fonction dont l'aspect dépend de la disposition des points.

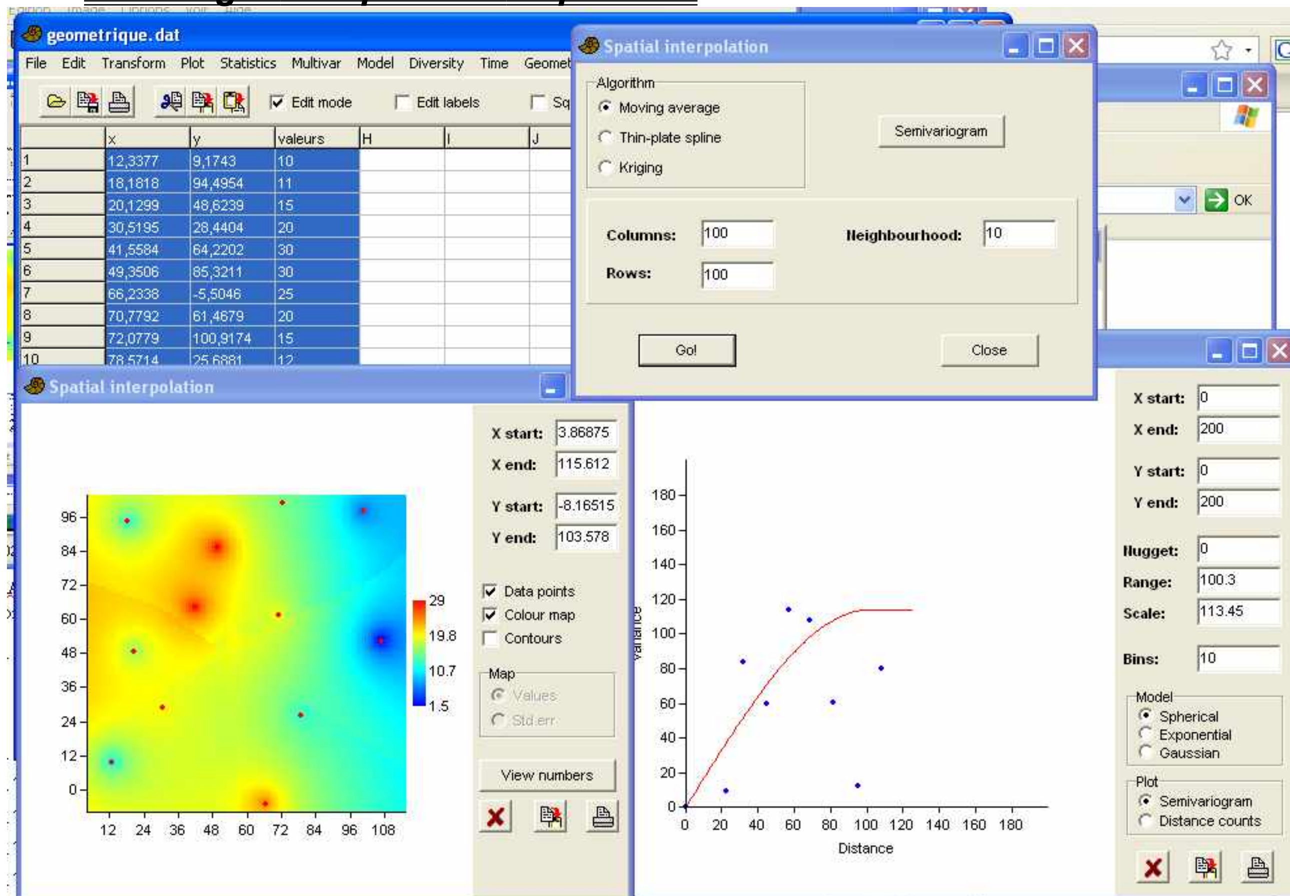
Si les points sont distribués au hasard, la fonction K(d), qui est la densité moyenne des points en fonction de la distance d, doit augmenter comme le carré de la distance.

La fonction L(d) vaut la racine carrée de K(d), et est donc proportionnelle à la distance, et L(d)-d doit être égale à zéro si les points sont bien aléatoires.

Par contre, si les points sont groupés en agrégats, les fonctions montrent des brisures.

« Area » indique la surface prise par l'ensemble des points. Par défaut, c'est le plus petit rectangle contenant les points, mais l'utilisateur peut changer cette valeur.

8.6 Gndding = interpolations spatiales



Cette fonction permet d'obtenir des cartes montrant l'estimation d'une grandeur pour toute la zone, à partir seulement de quelques points de mesures. C'est très spectaculaire, mais il ne faut pas accorder une confiance absolue dans les valeurs numériques des interpolations.

On doit fournir à PAST trois colonnes, correspondant aux divers points de mesure. La première valeur est l'abscisse x, la deuxième est l'ordonnée y, et la troisième est la valeur numérique de la grandeur observée.

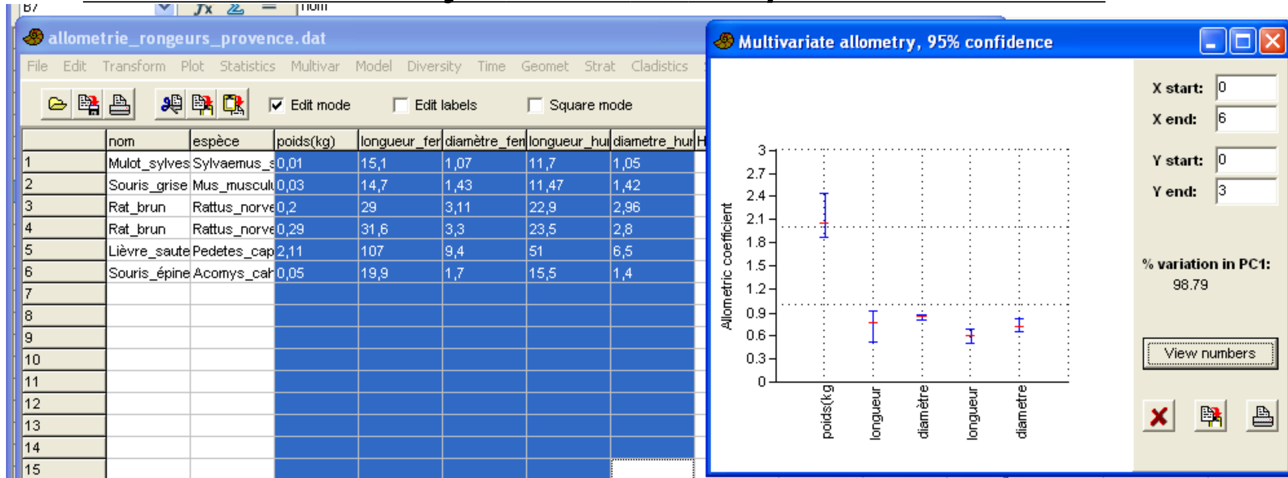
Le choix de l'option entraîne l'apparition d'une petite fenêtre. Par défaut, les calculs d'interpolation sont faits sur une matrice de 100 lignes et 100 colonnes, mais ces nombres peuvent être changés (plus la taille est grande, plus les calculs sont longs). « Neighbourhood: » indique le nombre de points pour le calcul de la moyenne mobile, et n'est valide que pour cette option, qui est l'option par défaut.

« Go » déclenche les calculs d'interpolation et le traçage de la carte.

On a le choix entre trois manières d'interpoler :

- « Moving average », par la méthode des moyennes mobiles, où chaque point a un poids en fonction inverse de la distance. C'est un algorithme simple, qui ne donne pas toujours de bons lissages, mais qui a l'avantage que les valeurs calculées ne sortent jamais de l'étendue des points de mesure.
- « Thin plate spline », qui donne le lissage maximal
- « Kriging », le krigeage, qui nécessite d'avoir auparavant calculé le semivariogramme.

8.7 Multivanate allometry = Allométrie sur plusieurs variables



On part d'un ensemble de colonnes correspondant aux grandeurs mesurées. PAST transforme ces données en logarithmes, puis réalise une analyse en composantes principales (ACP = PCA). La première composante principale est considérée comme l'axe de taille, et le coefficient d'allométrie est calculé pour chaque variable initiale, comme la contribution de cette variable à la première composante principale divisée par la contribution moyenne de toutes les variables.

8.8 Divers types d'analyses de forme (analyses procustéennes)

Le terme « procustéen » fait référence à la mythologie grecque, où le bandit Procuste (Procrustes en anglais) forçait ses victimes à s'allonger dans un lit, et les adaptait à la longueur de son lit, en les coupant à la hache ou en les étirant ; il fut finalement tué par Thésée. Ce terme « procustéen » fait maintenant référence aux problèmes d'ajustement de formes, par translation, rotation et changement d'échelle.

Diverses fonctions demandent que les données soient normalisées par la transformation appropriée (voir menu Transform | Procrustes (2D/3D)).

8.8.1 Fourier shape(2D)

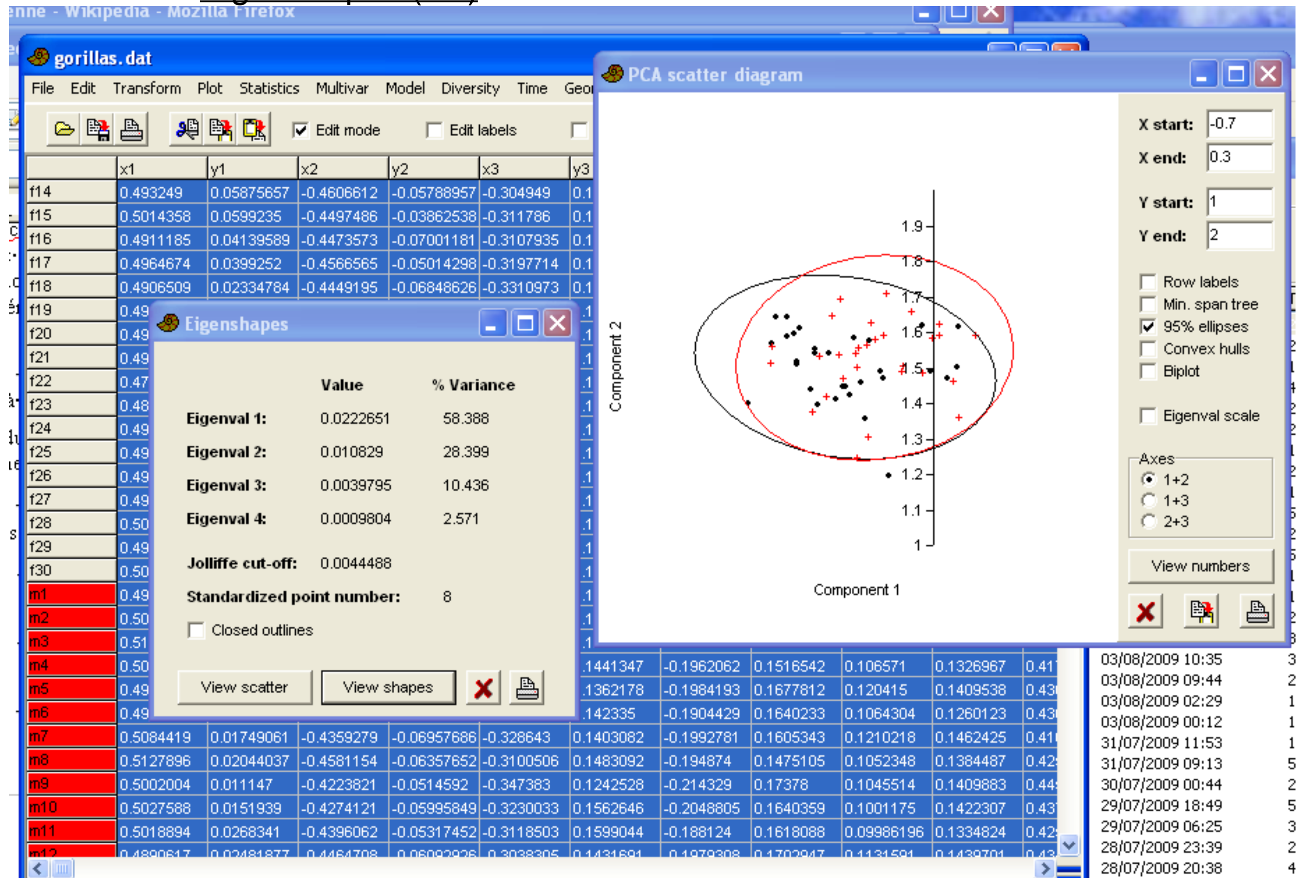
Il faut fournir à cette option au moins 6 colonnes, c'est à dire 3 paires de coordonnées (x,y).

Chaque individu mesuré correspond à une ligne, et les mesures correspondent à des couples x,y, disposés sur une même ligne.

8.8.2 Elliptic Fourier (2D)

Il faut au moins dix couples de points (x,y) sur une même ligne, donc vingt colonnes.

8.8.3 Eigenshapes (2D)



Le résultat est une sorte d'analyse en composantes principales (PCA = ACP) portant sur les positions des points indiqués.

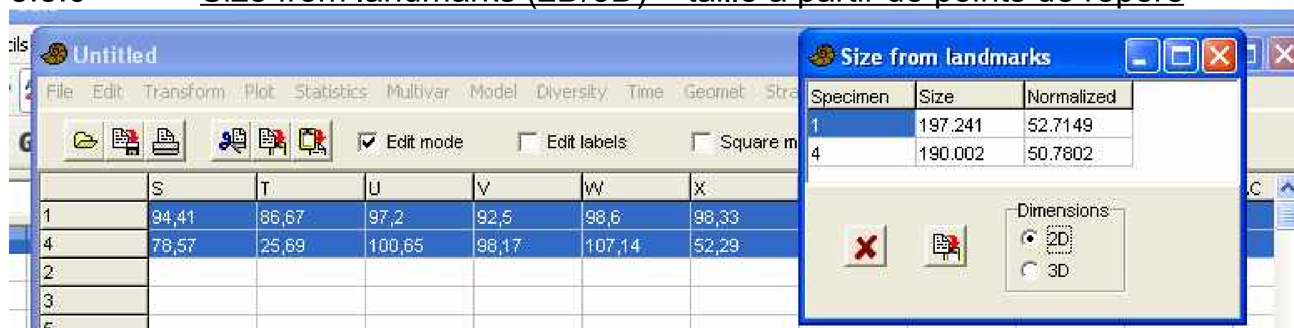
8.8.4 Thin-plate splines and warps (2D) = déformation et lissages

La référence prise est soit le premier spécimen (première ligne), soit la moyenne de l'ensemble des spécimens. Les déformations des autres spécimens sont calculées, et l'on passe d'un spécimen à un autre par en cliquant sur les boutons de navigation (spinbuttons : double triangle noir). La normalisation de Procuste (voir menu Transform | Procustes) est recommandée.

8.8.5 Relative warps (2D) = déformations relatives

Là aussi, la normalisation de Procuste est recommandée.

8.8.6 Size from landmarks (2D/3D) = taille à partir de points de repère



Cette option calcule la taille des échantillons (chaque échantillon correspond à une ligne). La

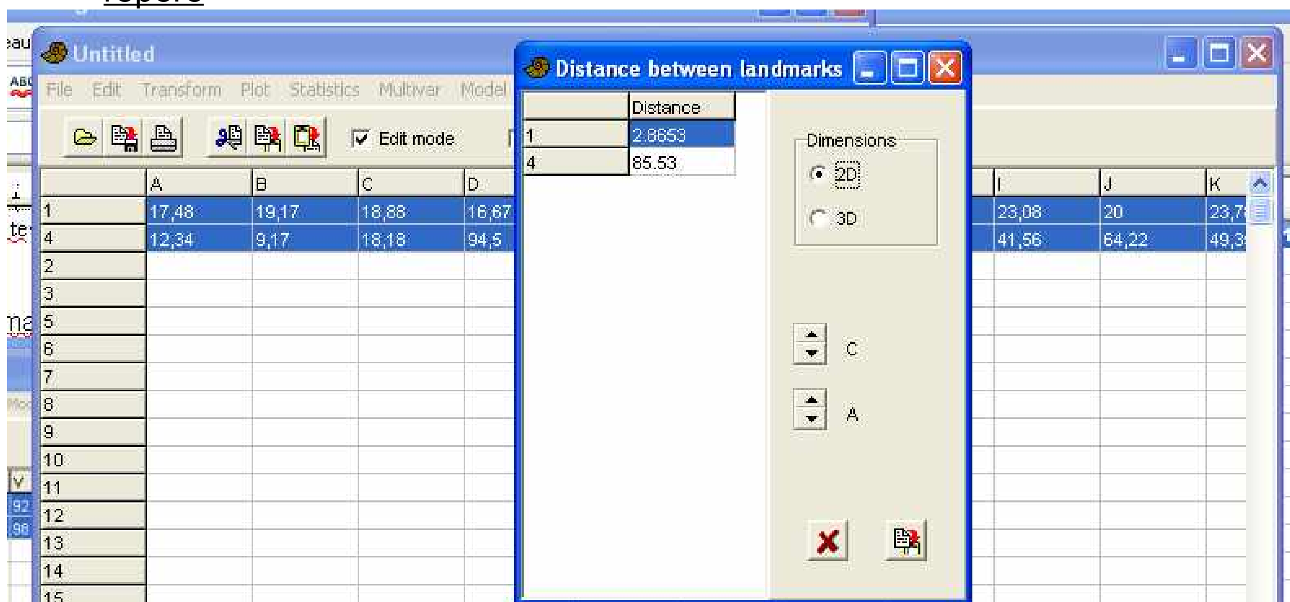
distance est la norme euclidienne des distances de tous les points de repères au centroïde.

Par défaut, PAST considère qu'il y a deux dimensions, et il faut que chaque ligne contienne un nombre pair de cases, car chaque couple de cases correspond à un couple x, y. On peut déclarer travailler en 3 dimensions en cochant la case « 3D », et dans ce cas, il faut que le nombre de colonne soit un multiple de 3.

« Normalized » correspond à la taille calculée précédemment divisée par la racine carrée du nombre de points, ce qui permet de comparer les tailles des spécimens ayant un nombre différent de points de repère.

Pour cette fonction, il ne faut pas effectuer de normalisation de Procuste.

8.8.7 Distance from landmarks (2D/3D) = distance à partir de points de repère



Cette fonction calcule les distances entre points de repère d'un même spécimen. L'exemple ci-dessus calcule pour les deux spécimens 1 et 4 la distance entre le point (C,D) et le point (A,B). La première distance vaut 2,8653, car c'est la racine de $(18,88-17,48)^2 + (16,67-19,17)^2$.

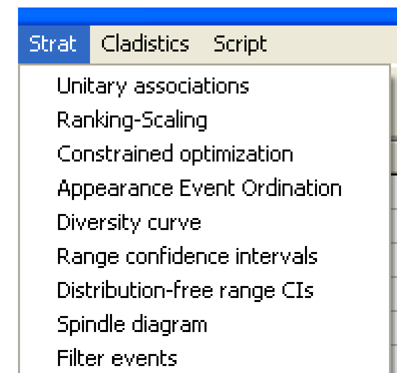
On peut naviguer dans les points dont il faut calculer la distance par les boutons de navigation. Ces boutons ont un pas de 2 colonnes en 2 dimensions, puisque chaque point correspond à un couple de colonnes.

8.8.8 All distances from landmarks (EDMA) = toutes les distances entre points de repère

Cette option transforme le tableau, en calculant toutes les distances entre points de repère d'un même spécimen.

8.8.9 Landmark linking = liens de points de repères

9 Strat : analyses stratigraphiques spécifiques



9.1 *Unitary associations = associations unitaires*

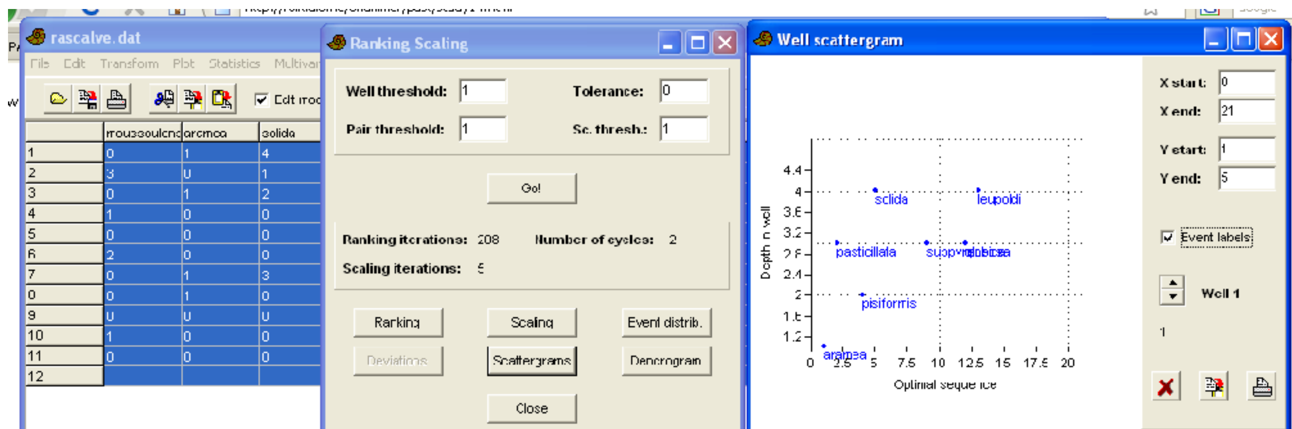
Il faut une matrice de présence/absence, avec les colonnes correspondant aux taxons. Les lignes correspondent aux observations, avec diverses contraintes :

– les différents lieux d'observation doivent être notés par des couleurs différentes. Ces couleurs permettent à PAST de faire la liste des lieux, mais la couleur en elle-même n'a pas d'importance ; on peut réutiliser plusieurs fois les mêmes couleurs, en particulier pour des tableaux comportant de nombreux lieux d'observation.

– Pour un même lieu, les observations doivent être triées chronologiquement (donc par position, en stratigraphie) : les strates les plus basses (les plus anciennes) doivent être les plus basses dans le tableau.

Après avoir choisi cette fonction « Unitary associations », puis cliqué sur « Go! », on peut choisir diverses possibilités pour mieux comprendre les relations entre les diverses strates des divers lieux.

9.2 Ranking-Scaling = classement des données stratigraphiques



Il faut partir d'une table de profondeur des événements dans divers puits ou lieux d'observation.

Les lignes indiquent les puits, les colonnes indiquent les événements, et la valeur dans les cases correspond à la profondeur où ces événements sont observés dans les puits (les absences sont codées par des 0). Si l'on n'a pas de mesure de profondeur, mais seulement un ordre d'apparition dans les puits, on peut réaliser le codage par des nombres entiers correspondant au rang.

La première étape est le classement (« ranking »), optimisé pour l'ensemble des puits, même s'il existe des contradictions entre les divers puits. Par exemple, il est possible que dans le puits 1, l'événement A soit avant l'événement B, alors que c'est le contraire dans le puits 2.

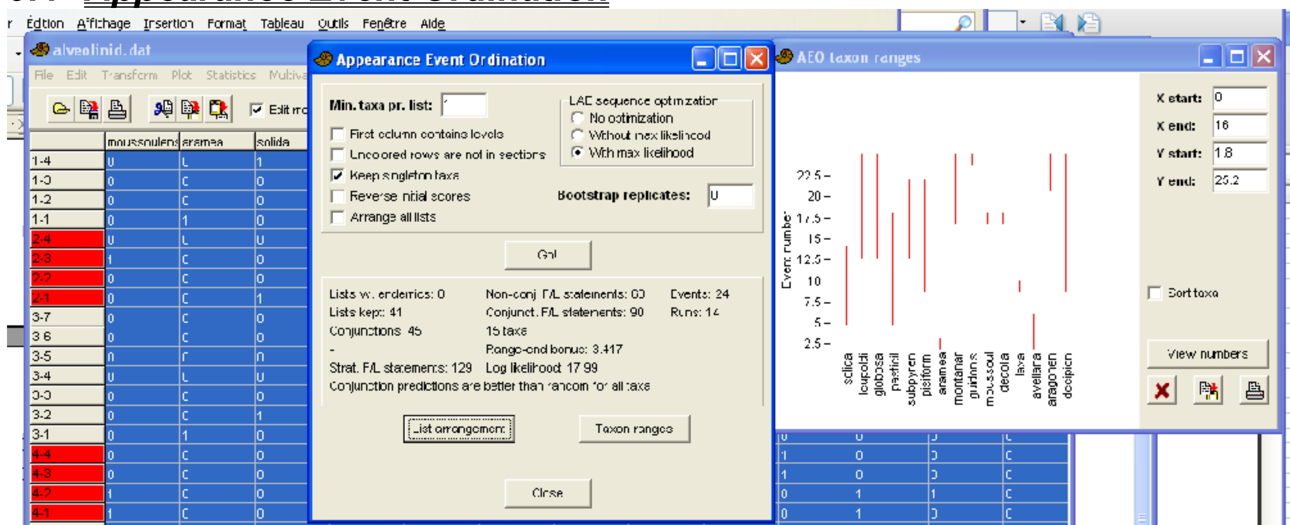
La deuxième étape est la mise à l'échelle (« scaling »). Cette étape refait aussi un classement.

9.3 Constrained optimization CONOP

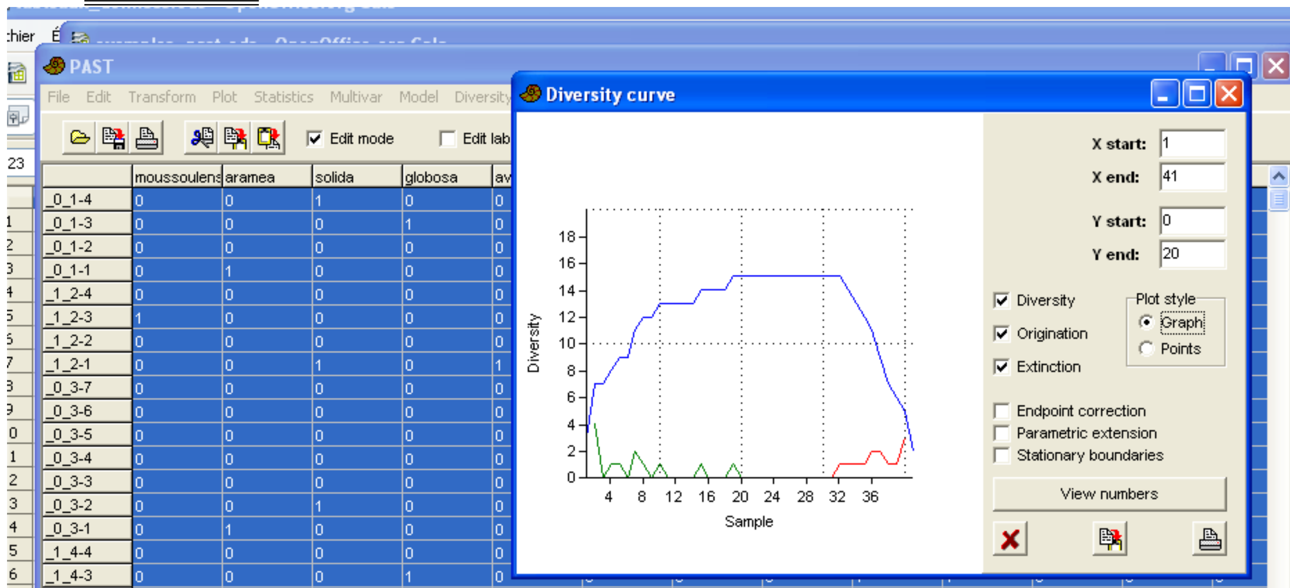
Là encore, les lieux d'observation (les puits d'un forage, par exemple) correspondent aux lignes. Dans les cases des colonnes impaires sont les profondeurs ou niveaux d'apparition du taxon, et dans les colonnes paires les profondeurs de disparition du taxon. Les événements manquants sont codés par des zéros.

Cette fonction permet de faire des corrélations biostratigraphiques quantitatives.

9.4 Appearance Event Ordination



9.5 Diversity curve = courbe de diversité, avec apparition et disparition de taxons



A partir d'une matrice de présence-absence, où les colonnes correspondent aux espèces et les lignes aux étages, PAST trace un diagramme avec en bleu le nombre d'espèces présentes. En cochant « Origination », il apparaît une courbe verte indiquant les apparitions d'espèces, et en cochant « Extinction », il apparaît une courbe rouge montrant les extinctions.

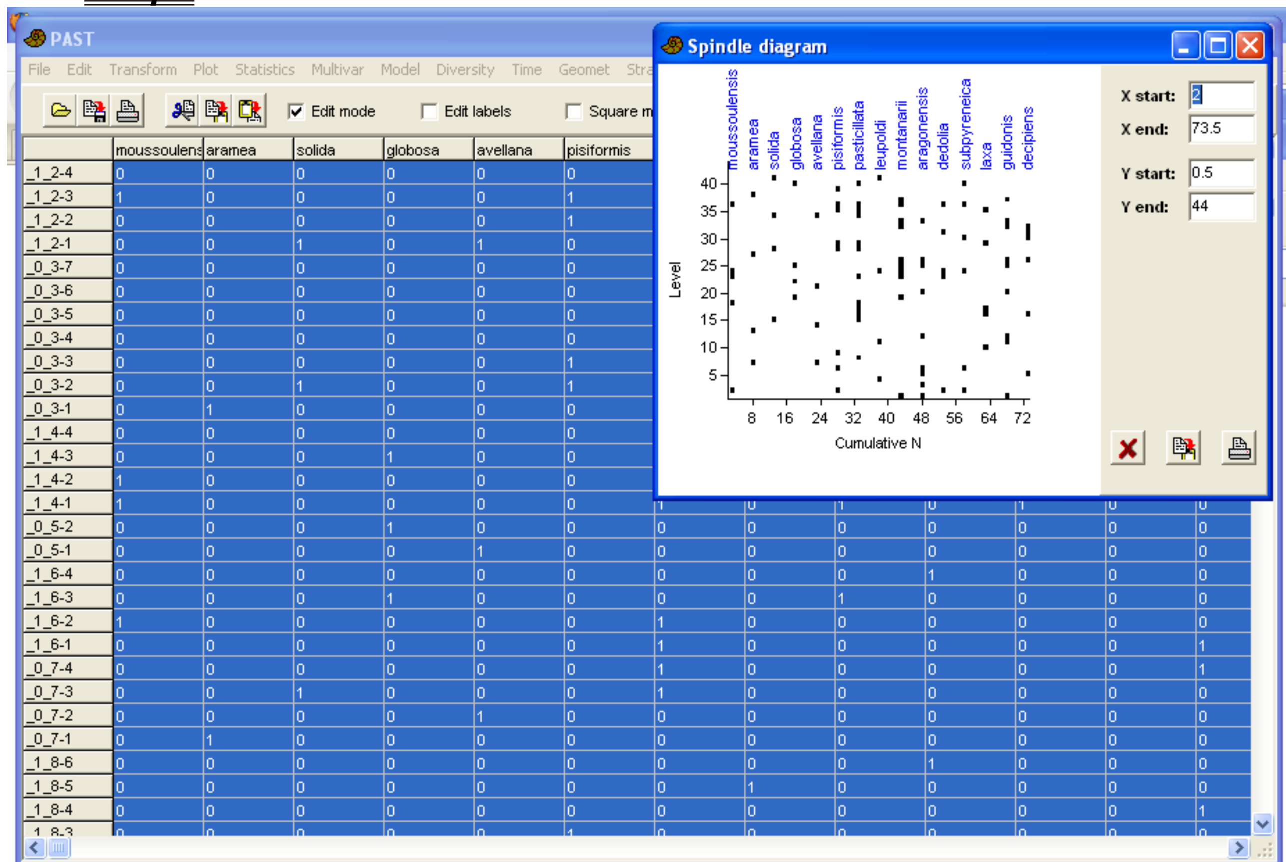
9.6 (Range confidence intervals = calcul des intervalles de confiance de la distribution des fossiles)

(calculs théoriques, ne dépendant pas des données dans le tableau. On suppose une distribution aléatoire des fossiles dans les horizons)

9.7 Distribution-free range CIs = intervalles de confiance indépendants de la distribution.

Ici, on ne suppose pas une distribution aléatoire des horizons fossilifères. On suppose qu'il n'y a pas de corrélation entre la position stratigraphique et la taille de l'intervalle.

9.8 *Spindle diagram = graphique d'observation des taxons au cours du temps*



Cette option fait simplement un graphique par points à partir d'une grosse matrice de présence/absence. Elle peut être utile pour visualiser une très grosse matrice.

9.9 *Filter events*

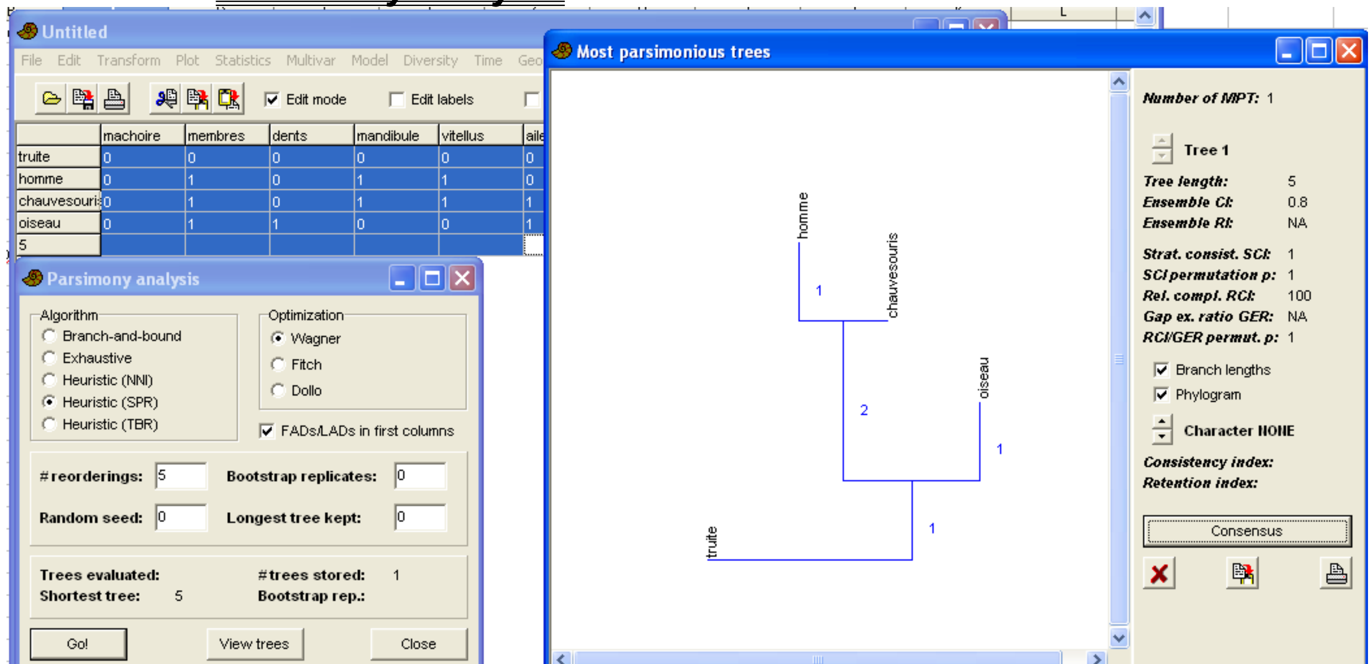
(transformation des données)

10 Cladistics : analyses cladistiques

Il n'y a qu'une seule option, l'analyse d'une matrice de caractères par parcimonie.

Ce n'est pas la seule méthode d'utiliser de telles matrices pour imaginer un arbre phylogénétique. On peut aussi utiliser l'analyse multivariée, en particulier le regroupement (menu « Multivar | cluster analysis »), mais les différentes méthodes de calcul de distance ne donnent pas le même résultat.

10.1 Parsimony analysis



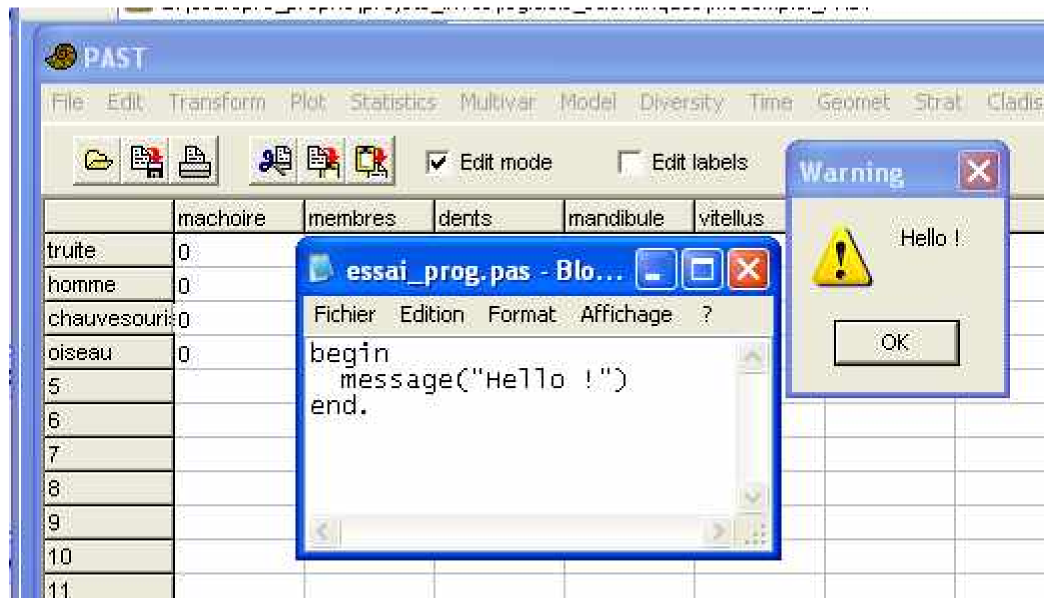
Les divers taxons correspondent aux lignes, et les divers caractères correspondent aux colonnes.

L'extragroupe doit être mis en première ligne.

Les variantes des caractères doivent être codés par des entiers entre 0 et 255.

11 Script : programmation des actions de PAST

PAST contient un embryon de fonctions permettant sa programmation. Le langage peut être considéré comme une variante du langage Pascal, mais très simplifié. Normalement, cette partie « programmation » devrait progresser dans les versions ultérieures de PAST.



11.1 Chargement et exécution d'un programme

Le programme lui-même est un simple fichier de texte, à rédiger à l'aide d'un éditeur de texte.

11.1.1 Load script

Cette fonction charge en mémoire un programme, et rend possible l'option suivante.

11.1.2 Run

Cette fonction exécute un programme préalablement chargé en mémoire.

11.2 Structure du langage

Les variables peuvent avoir des nom de longueur quelconque, et ne doivent pas être déclarées avant leur emploi (déclaration implicite au moment du premier emploi). Ces variables ont toutes une portée globale. L'opérateur d'affectation est :=.

On peut mettre des commentaires : tout ce qui suit le dièse # est considéré comme un commentaire.

Il existe quatre types de variables : des nombres (réels en double précision), des chaînes de caractères, des vecteurs et des tableaux. Ces types sont implicitement définis au premier usage de la variable, et les conversions de type sont faites automatiquement.

L'ensemble du script doit être encadré par un couple begin...end, et les blocs à l'intérieur du programme sont aussi encadrés par des couples begin … end.

Les structures de contrôle sont d'une part les instructions conditionnelles par if ….begin … end else …, d'autre part les boucles for et while.

Un prédicat « vrai » est traduit par 1, et un prédicat « faux » est traduit par 0.

Il est possible de structurer le programme par des procédures, mais qui sont beaucoup plus simples

que dans le langage Pascal normal : elles n'ont pas de paramètres, pas de valeur de retour, pas de variables locales. Elles doivent être définies après le begin de début du programme. Bizarrement, l'appel de la procédure doit se faire avec un paramètre quelconque (fictif), alors que la procédure avait été déclarée sans paramètres.

11.2.1 Opérateurs

Les opérateurs mathématiques classiques +, -, *, /, ^ existent.

La concaténation de chaînes est faite par &.

Les opérateurs de comparaison sont =, >, <, >=, <=, <>.

11.2.2 Fonctions diverses

On trouve les fonctions mathématiques classiques cos, sin, tan, exp, log, sqrt, int, rnd, mais aussi des fonctions statistiques moins courantes.

Diverses opérations sont possibles sur les vecteurs et matrices, en particulier l'inversion, le calcul de moyenne, le calcul de covariance, le classement... Diverses opérations de calcul de distances entre matrices sont possibles.

On peut aussi programme le traçage de graphique par des instructions de type drawline, drawstring, drawhistogram...

Bibliographie et références

L'article officiel décrivant PAST :
Hammer, Ø., Harper, D.A.T., and P. D. Ryan, 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica 4(1): 9pp.

Quelques livres de statistiques naturalistes :
Ancelle Thierry « Satistique épidémiologie » Maloine 2002

Bouyer Jean « Méthodes statistiques Médecine-Biologie » INSERM 1997

Geller S. « Abrégé de statistique » Masson 1979

Laberche Jean-Claude « Statistiques et expérimentation en biologie » Ellipses 2008

Quelques sites internet :
Définitions et mesure de la biodiversité
http://www.agroparistech.fr/IMG/pdf/Biodiv2008.pdf

LES PRINCIPAUX OUTILS STATISTIQUES EN ECOLOGIE
http://alexandra.gigou.googlepages.com/Rapportcologienumrique.pdf

Synthèse des méthodes d'évaluation de la qualité du benthos en milieu côtier
http://www.rebent.org/documents/document.php?g_id_document=166

Index